QUEUEING MODELS FOR LARGE SCALE CALL CENTERS

A Thesis
Presented to
The Academic Faculty

by

Joshua E Reed

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2007

QUEUEING MODELS FOR LARGE SCALE CALL CENTERS

Approved by:

Jim Dai, Committee Co-Chair
School of Industrial and Systems Engineering
Georgia Institute of Technology

Amy Ward, Committee Co-Chair
School of Industrial and Systems Engineering
Georgia Institute of Technology

Ronald Billings
School of Industrial and Systems Engineering
Georgia Institute of Technology

Robert Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology

Marty Reiman
Bell Labs

Date Approved: 15 May 2007

To my parents and grandparents...

ACKNOWLEDGEMENTS

I am indebted to many people for their help and support during my graduate studies. This thesis would not have been possible without them. First and foremost, I would like to thank my advisors Jim Dai and Amy Ward.

I would like to thank Dr. Ward for serving as my principal advisor for the first half of my Ph.D. studies. Dr. Ward provided me with a solid foundation in the field of heavy traffic theory and I am thankful to her for the devotion and support which she showed me as her student.

I would next like to thank Dr. Dai for assuming responsibility as my principal advisor for the second half of my studies. I am especially thankful to him for his insistence on always finding the proper and most elegant proof of a result. This is something which I will always carry with me as I continue on in my work.

I would also like to thank Marty Reiman, Bob Foley and Ronald Billings for agreeing to serve on my thesis committee. I would especially like to thank Marty Reiman for taking the time to fly to Atlanta for my defense.

The head of our graduate studies program, Dr. Gary Parker, performed many small miracles for me over the past several years without which I would not have completed this thesis and I will always be grateful to him for the encouragement he has given me.

The ARCS Foundation deserves special mention for their generous financial support throughout my graduate studies.

There are many other Professors who have been very influential in my thesis work. I would like to thank Erica Plambeck for hiring me as her research assistant for a year.

I would also like to thank Ward Whitt for his extremely helpful comments and suggestions on Chapter II of this thesis. His work has provided much of the inspiration for this thesis.

I am also very appreciative of the unwavering support my friends and family back home

in Miami had for me throughout this work.

Finally, I would like to thank my friends, without whom this work would not have been nearly as enjoyable. There are so many of them to thank that I cannot list them all. I will simply say that the details of many of them could perhaps fill up yet another volume of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY

The call center industry is rapidly expanding. One of the most prominent features of a modern call center is that it may employ hundreds if not thousands of agents. Recently, many researchers have studied queueing models for such large scale call centers in a many-server heavy traffic regime know as the Halfin-Whitt [15] regime. The Halfin-Whitt regime is achieved by considering a sequence of $M/M/N$ queues where the arrival rate and the number of servers grows to infinity while the probability of delay is held fixed at a number strictly between zero and one.

In the first half of this thesis, we extend the results of Halfin and Whitt to generally distributed service times. This is accomplished by first writing the system equations for the $G/GI/N$ queue in a manner similar to the system equations for $G/GI/\infty$ queue. We next identify a key relationship between these two queues. This relationship allows us to leverage several existing results for the $G/GI/\infty$ queue in order to prove our main result. Our main result in the first part of this thesis is to show that the diffusion scaled queue length process for the $G/GI/N$ queue in the Halfin-Whitt regime converges to a limiting stochastic process which is driven by a Gaussian process and satisfies a stochastic convolution equation. We also show that a similar result holds true for the fluid scaled queue length process under general initial conditions.

Customer abandonment is also a common feature of many call centers. Some researchers have even gone so far as to suggest that the level of customer abandonment is the single most important metric with regards to a call center's performance. In the second half of this thesis, we improve upon a result of Ward and Glynn's [52] concerning the $GI/GI/1+GI$ queue in heavy traffic. Whereas Ward and Glynn obtain a diffusion limit result for the $GI/GI/1+GI$ queue in heavy traffic which incorporates only the density the abandonment distribution at the origin, our result incorporate the entire abandonment distribution. This is accomplished by a scaling the hazard rate function of the abandonment distribution as

the system moves into heavy traffic. Our main results are to obtain diffusion limits for the properly scaled workload and queue length processes in the $GI/GI/1+GI$ queue. The limiting diffusions we obtain are reflected at the origin with a negative drift which is dependent upon the hazard rate of the abandonment distribution. Because these diffusions have an analytically tractable steady-state distribution, they can be used to provide a closed-form approximation for the steady-state distribution of the queue length and workload processes in a $GI/GI/1+GI$ queue. We demonstrate the accuracy of these approximations through simulation.

CHAPTER I

INTRODUCTION

The call center industry has experienced phenomenal growth over the past decade. In the United States alone, there are an estimated 70,000 call centers and year over year growth has been steady at a rate of over 20% a year for the past several years. Annual expenditures on call centers range from anywhere between $100 to $300 billion [48] and it is estimated that somewhere between 50-75% of a call center's costs are labor related. According the U.S. Bureau of Labor Statistics [1], the call center industry accounts for more than 1.4% of private sector employment, with the total number of agents in the U.S. residing at approximately 1.55 million.

The most widely used model in the management of call centers is the $M/M/N$ queue which also known as Erlang C [13]. This model assumes Poisson arrivals and exponentially distributed service times. Its popularity is due at least in part to the fact that it provides closed form, tractable formulas for several steady state quantities of interest, such as the long run average probability that a customer will have to wait and the long run average queue length.

Consider an $M/M/N$ queue with arrival rate λ and service rate μ and denote by $Q(t)$ the total number of customers in the system at time t. It is well known [33] that $Q = \{Q(t), t \geq 0\}$ is a continuous time Markov chain with state space $\mathbb{N} = \{0, 1, 2, ...\}$. It is also positive recurrent if and only if $\rho = \lambda/(N\mu) < 1$. In this case, we may solve its balance equations in order to determine its steady state distribution. In particular, the distribution of the steady state number of customers in the system is given by $P(Q(\infty) = i) = p_0\rho_i,\ i \geq 0$, where

$$\rho_i = \begin{cases} \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i & \text{if } 0 \leq i \leq N-1, \\ \frac{N^N}{N!}\rho^i & \text{if } i \geq N \end{cases} \tag{1}$$

and

$$p_0 \;=\; \left(\sum_{i=0}^{N-1} \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i + \frac{N^N}{N!}\cdot\frac{\rho^N}{1-\rho}\right)^{-1}. \tag{2}$$

The steady state probability that a customer will have to wait before being served may also be calculated and is given by

$$\alpha \;=\; P(Q(\infty)\geq N) \;=\; \frac{N^N}{N!}\cdot\frac{\rho^N}{(1-\rho)}\cdot p_0, \tag{3}$$

where p_0 is given by (2). The quantity α in (3) is also commonly referred to as the Erlang delay formula.

Unfortunately, the formulas in (1) and (3) may be hard to compute when either the number of servers N is large or the traffic intensity of the system $\rho = \lambda/(N\mu)$ is close to one. Moreover, (1) and (3) do not seem to yield any practical insights into how the performance of the system becomes affected as its capacity is adjusted. This is an especially important point for call center managers who would like to know exactly how many agents they should hire in order to meet a predetermined level of service.

One approach which had yielded great success in providing analytical approximations to formulas such as (1) and (3) is known as the heavy traffic approach. Originally developed by Kingman [26, 27] for the single server queue, the approach yields close approximations to (1) and (3) when the traffic intensity of the system of the close to one. It also gives approximations to the quantities in (1) and (3) for generally distributed interarrival and service times as well. Consider a sequence of $GI/GI/1$ queues indexed by r, where the interarrival times to the r^{th} queue are given by the mean 1, i.i.d. sequence $\{u_i, i\geq 1\}$ and the service time sequence is given by $\{\rho^r v_i, i\geq 1\}$ where $\{v_i, i\geq 1\}$ is an i.i.d. sequence of mean 1 random variables. It is clear that in this case, the traffic intensity of the r^{th} system is given by $\rho^r < 1$. Furthermore, setting $\tilde{Q}^r(\infty) = (1-\rho^r)Q^r(\infty)$, one then has the following result.

Theorem (Kingman '61). *If $\rho^r \uparrow 1$ as $r\to\infty$, then $\tilde{Q}^r(\infty) \Rightarrow \tilde{Q}(\infty)$ as $r\to\infty$ where $\tilde{Q}(\infty) \sim exp(2/(var(u_1)+var(v_1)))$.*

The seminal work of Kingman has come to form the basis of what is now known as conventional heavy traffic theory. This theory has had spectacular success in solving problems

in such diverse fields as inventory control and manufacturing and logistics. However, conventional heavy traffic theory is not well suited to the case in which the number of servers is large. Furthermore, it does not incorporate customer abandonment. Both of these features become important when modeling call centers.

A large scale call center may employ hundreds if not thousands of agents. In this case, the conventional heavy traffic regime no longer applies. There is, fortunately, however, an alternative "unconventional" heavy traffic which is very well suited for problems with large numbers of servers. This regime was originally discovered by Halfin and Whitt in their seminal work [18] and has now come to be known as the Halfin-Whitt regime.

Consider a sequence of $M/M/N$ queues indexed by r where the r^{th} system has N^r servers. Let λ^r denote the arrival rate to the r^{th} system and assume that $\lambda^r \to \infty$ as $r \to \infty$. Assume further that the service rate is held fixed at 1 for all r and define $\rho^r = \lambda^r/(N^r\mu) < 1$ to be the traffic intensity of the r^{th} system. Let $Q^r(\infty)$ be the steady queue length of the r^{th} system and define $\tilde{Q}^r(\infty) = (N^r)^{-1/2}(Q^r(\infty) - N^r)$. If we now let Φ be the standard normal cumulative distribution function and define

$$H(x) \quad = \quad (1 + \sqrt{2\pi}x\Phi(x)e^{x^2/2})^{-1}, \quad x \geq 0,$$

then, Theorem 1 of Haflin and Whitt [15] may be stated as follows.

Theorem (Halfin and Whitt '81). *If $N^r(1-\rho^r) \to \beta$ as $r \to \infty$ where $\beta > 0$, then $\tilde{Q}^r(\infty) \Rightarrow \tilde{Q}(\infty)$ as $r \to \infty$, where $P(\tilde{Q}(\infty) \geq 0) = H(\beta)$, $P(\tilde{Q}(\infty) > x|\tilde{Q}(\infty) > 0) = e^{-x\beta}, x \geq 0$, and $P(\tilde{Q}(\infty) \leq x|\tilde{Q}(\infty) \leq 0) = \Phi(\beta + x)/\Phi(\beta), x \leq 0$.*

Halfin and Whitt's Theorem above provides approximations to (1) and (3) for the case in which N is at least moderately large. It is for this reason that the Halfin-Whitt regime is also commonly refereed to as the many-server regime. There is, however, one significant difference between Halfin and Whitt's result and that of Kingman's. While Kingman's result holds for generally distributed interarrival and service times, Halfin and Whitt's results have historically required the assumption of exponentially distributed service times. This may be viewed as a limitation from the perspective of call center applications as it is now commonly accepted that service times at call centers are not in general exponentially distributed [7].

Recently, however, several authors have begun to obtain Halfin-Whitt type convergence results for more general classes of service time distributions. Puhalskii and Reiman [40] have demonstrated convergence of the $G/PH/N$ queue length process in the Halfin-Whitt regime, where PH stands for phase type service time distributions. Their approach is to consider a multidimensional Markovian process where each dimension corresponds to a different phase of the service time distribution. Jelenković, Mandelbaum, and Momčilović [22] have shown convergence of the steady state distribution of the $GI/D/N$ queue, where D stands for deterministic service times. Their proof involves focusing on a single server in the system and studying its queue length behavior as it evolves over time. Whitt in [57] has shown process level convergence of the $G/H_2^*/N/M$ queue, where H_2^* stands for a mixture of an exponential random variable and a point mass at zero. In [37], Mandelbaum and Momčilović study the virtual waiting time process of $G/GI/N$ in the Halfin-Whitt regime assuming that the service time distribution possess finite support. Their approach relies on a combination of combinatorial and probabilistic arguments. Nevertheless, despite their successes, it does not appear that any of the aforementioned approaches may easily be extended to the case of general service time distributions and so to this date there has remained no general methodology for analyzing the $G/GI/N$ queue in the Halfin-Whitt regime. This is the main contribution of the first half of this thesis.

The main insight to our approach is to write the system equations in a manner similar to the system equations for the $G/GI/\infty$ queue. Proposition 2.1.1 in Section 2.1.2 then provides a crucial link between our system and the $G/GI/\infty$ queue which allows the asymptotic analysis to proceed. In an effort to give a quick idea of what our main results, first recall the heavy traffic results of Borvokov [6] and Krichagina and Puhalskii [28] for the $G/GI/\infty$ queue. Recall that in [6] and [28], heavy traffic for the $G/GI/\infty$ queue is defined by letting the arrival rate to the system grow large while holding the service time distribution fixed. In such a regime, it can be shown that the properly centered and scaled queue length processes will converge to a Gaussian process. Let us therefore denote by $\tilde{Q}_I$ the limiting Gaussian process obtained for a $G/GI/\infty$ queue with the same sequence of arrival processes and an identical service time distribution as in our original sequence of $G/GI/N$

queues. Then, the limiting process of Theorem 2.4.4 of Section 2.4 for the properly centered and scaled queue length process in our original sequence of $G/GI/N$ queues is given by the unique strong solution to

$$\tilde{Q}(t) \quad = \quad \tilde{M}_Q(t) + \tilde{Q}_I(t) + \int_0^t \tilde{Q}^+(t-s)dF(s), \quad \text{for} \quad t \geq 0, \tag{4}$$

where $\tilde{Q}^+ = \max(\tilde{Q}, 0)$, F is the CDF of the service time distribution and $\tilde{M}_Q$ is an additional process which is related to the initial conditions of the queue. Note that the additional integral term on the righthand side of (4) is naturally positive as one would expect more customers in a $G/GI/N$ queue than in a corresponding $G/GI/\infty$ queue. Corollary 2.4.5 in Section 2.4 also shows that (4) may be equivalently expressed as

$$\tilde{Q}(t) \quad = \quad \zeta(t) + \int_0^t \zeta(t-s)dM(s) - \int_0^t \tilde{Q}^-(t-s)dM(s), \quad \text{for} \quad t \geq 0, \tag{5}$$

where $\zeta = \tilde{M}_Q + \tilde{Q}_I$, $\tilde{Q}^- = -\min(0, \tilde{Q})$ and M is the renewal function associated with the pure renewal process with interarrival distribution F. From (5), it is then a matter of a few direct calculations to recover Halfin and Whitt's original results. We should also point out that along the way we develop fluid limit results which closely resemble (4) above.

In the second half of this thesis, we consider customer abandonment. Although many standard queueing models assume that customers are infinitely patient while waiting for service, it is often the case that such an assumption is not reasonable. In particular, impatient customers faced with long waiting times often evidence their frustration by abandoning the system before completing service. For example, call center callers placed on hold frequently hang up while waiting for an agent to assist them, and web browser users often cancel their viewing requests in the face of long download times.

To the best of our knowledge, the first person to remark on the importance of incorporating customer abandonment in a queueing model was Palm [39], who witnessed the impatient behavior of telephone switchboard customers. More recent studies have established that Markovian M/M/n+M abandonment models (where the final +M represents the abandonment distribution) are well-approximated by diffusion processes in heavy traffic, both when the number of servers grows large (see Garnett et al [14]), and when the number of servers remains fixed (see Ward and Glynn [50]). However, in reality, it is often

the case that customer abandonment times are not exponentially distributed, as exhibited by the study of Brown et al [7] of a bank call center data set. Although stability results for models in very general frameworks exist (see, for example, Stanford [47], Baccelli, Boyer, and Hebuterne [3], Lillo and Martin [35] and Bambos and Ward [49]), the problem of rigorously establishing diffusion approximations for systems with general abandonment time distributions still remains an area of active research. Zeltyn and Mandelbaum [59] consider many server systems with generally distributed abandonment times, and Ward and Glynn [52] consider a single-server system with generally distributed abandonment times. Both works develop a heavy traffic asymptotic regime in which the limiting diffusion depends on the abandonment distribution only through the value of its density at 0.

The value of the density of a distribution at a single point is not a very robust statistic. For example, the estimated hazard rate function associated with the abandonment times of a U.S. bank's call center customers displayed in Figure 12 in the Internet Supplement to Zeltyn and Mandelbaum [59] is unstable near the origin. (Note that because abandonment times are non-negative, the values of both the hazard rate function and density associated with the abandonment distribution coincide at the point 0.) Therefore, identifying a limiting regime that preserves more of the structure of the abandonment distribution is of interest.

The main contribution of the second half of this thesis is to rigorously establish a heavy traffic regime for a single server queue operating under the FIFO service discipline with renewal arrivals, general service times, and general abandonment times in which the *entire* abandonment time distribution appears in the limiting diffusion approximation. We study both unbounded and bounded abandonment distributions, and develop diffusion approximations for both the offered waiting time process (the process that tracks the amount of time an infinitely patient arriving customer would wait for service) and the queue-length process. In so doing, we also prove results on the existence, uniqueness, and continuity of generalizations of the one-sided regulator mapping introduced in Skorokhod [46] and the two-sided regulator mapping having an explicit formula given in Kruk et al [29] to allow for a general, non-linear state dependence. Our key insight is to model customer abandonment times using the hazard rate function associated with the assumed abandonment time

distribution, as suggested by Whitt [56].

To specify our proposed diffusion approximation for the offered waiting time and queue-length processes (for simplicity we assume mean service times are one so that the proposed approximations for these two processes are identical), consider a single-server, FIFO queue having renewal arrival and service processes with identical rates n[1], in which each customer independently abandons the system if his service has not begun within an amount of time having a distribution with hazard rate function h and cumulative hazard function $H \equiv \int_0^x h(y)dy$. Let

$$H_D^n(x) \equiv \int_0^x h\left(\frac{y}{\sqrt{n}}\right) dy = \sqrt{n} H\left(\frac{x}{\sqrt{n}}\right). \tag{6}$$

Then, our suggested diffusion approximation for the scaled queue-length process $n^{-1/2}Q^n(\cdot)$ when n is large has infinitesimal drift $-H_D^n(x)$, where x is the state of the diffusion, constant infinitesimal variance that depends on the variance of the inter-arrival and service times, and is instantaneously reflected at the origin. When the distribution of abandonment times is bounded, our suggested approximating diffusion also has an upper reflecting barrier.

Table 1 displays the results of using our approximation to estimate the mean queue-length and abandonment probability in a queue with Poisson arrivals having rate $n = 2500$, deterministic service times having mean $1/2500$, and abandonment times distributed according to $G(p)$, a mean 1 gamma distribution having both scale and shape parameters equalling p. The cumulative hazard function associated with $G(p)$ is

$$H(x) \equiv -\ln\left(1 - \frac{\Gamma_{px}(p)}{\Gamma(p)}\right),$$

where $\Gamma(p) \equiv \int_0^\infty t^{p-1}e^{-t}dt$ is the gamma function, and $\Gamma_x(p) \equiv \int_0^x t^{p-1}e^{-t}dt$ is the incomplete gamma function ($p > 0$). From (6), the drift of our suggested approximating diffusion is

$$-H_D^n(x) = \sqrt{n}\ln\left(1 - \frac{\Gamma_{px/\sqrt{n}}(p)}{\Gamma(p)}\right). \tag{7}$$

Its variance is 1, which follows from Theorem 3.5.2, and its steady-state distribution is given in part (i) of Proposition 3.5.3. We ran each simulation to 2,000 time units so that

[1]Our analysis does not use the assumption of perfect balance; however, having equal arrival and service rates eases the exposition in the Introduction.

Table 1: A comparison of the simulated mean queue-length and abandonment probability for a GI/GI/1-GI queue with Poisson arrivals at rate 2500 per unit, deterministic service with mean 1/2500, and abandonment times distributed according to a gamma distribution with scale and shape parameter p.

	E[queue-length]			P[abandon]		
p	Simulated	Approximated	% Error	Simulated	Approximated	% Error
0.5	9.0093	8.418	6.57%	0.041292	0.043202	4.63%
2	84.911	86.835	2.27%	0.003367	0.003273	2.80%

the queue saw approximately 5,000,000 arrivals, and recorded the time average queue-length and abandonment fraction. Observe that all of our approximations in Table 1 differ from their simulated values by no more than 7%. In Chapter 3, we will present more extensive numerical results.

The remainder of this thesis will now be organized as follows. In the following subsection, we present the notation which will be used for the remainder of this work. In Chapter 2 we provide our results for the $G/GI/N$ queue in the Halfin-Whitt regime. Next, in Chapter 3, we present results for the $GI/GI/1+GI$ queue in a novel heavy traffic regime which will be defined there as well. In Chapter 4, we provide some concluding remarks. Proofs of some of the more technical results of Chapters 2 and 3 have been deferred to the appendices.

1.1 Notation

All random variables and processes are henceforth assumed to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each positive integer d, we let $D([0,\infty), \mathbb{R}^d)$ be the space of right continuous functions with left limits (RCLL) in $\mathbb{R}^d$ having time domain $[0,\infty)$. For convenience, we will use $D[0,\infty)$ to denote $D([0,\infty), \mathbb{R})$. We endow $D([0,\infty), \mathbb{R}^d)$ with the usual Skorokhod J_1 topology and let M^d denote the Borel σ-algebra associated with the J_1 topology. Stochastic processes are assumed to be measurable maps from $(\Omega, \mathcal{F})$ to $(D([0,\infty), \mathbb{R}^d), M^d)$ for some appropriate dimension d. A sequence of functions $\{x^n\} \in D([0,\infty), \mathbb{R}^d)$ is said to converge uniformly on compact sets to $x \in D([0,\infty), \mathbb{R}^d)$, if for each $T > 0$,

$$\max_{i=1,\ldots,d} \sup_{0 \le t \le T} |x_i^n(t) - x_i(t)| \to 0 \text{ as } n \to \infty.$$

The notation $\Rightarrow$ will be used to denote convergence in distribution and the notation $\stackrel{D}{=}$ means equal in distribution. When writing Lesbegue-Stieltjes integrals, we denote by $\int_a^b$ the integral over the closed interval $[a, b]$ and by $\int_{a+}^b$ the integral over the half-open interval $(a, b]$.

CHAPTER II

THE G/GI/N QUEUE IN THE HALFIN-WHITT REGIME

We extend the pioneering results of Halfin and Whitt [15] on the $GI/M/N$ queue to the more general $G/GI/N$ queue. This work is to a large extent motivated by the desire to better understand the performance of call centers which operate on a large scale. In recent years, the Halfin-Whitt regime, also known as the Q.E.D. regime, has been proposed as a suitable setting in which call centers may be analyzed. This is mainly for two reasons. First, in the Halfin-Whitt regime it is assumed that the arrival rate to the system is large and second, it is also assumed that the traffic intensity is close to one. It is natural to expect that in a large and efficiently run call center, both of these conditions will prevail. Unfortunately, however, results of Halfin and Whitt's type have mainly been limited to the assumption of exponentially distributed service times while more recent studies such as those by Brown et. al [7] seem to suggest that, at least in certain cases, service times at call centers may be lognormally distributed. It is therefore necessary to provide a suitable framework in which Halfin and Whitt's results may be extended to general service time distributions; this is the topic of the present chapter.

2.1 Model Formulation

2.1.1 System Equations

In this subsection, we will describe the system equation for the $G/GI/N$ queue. One of the key insights from Halfin and Whitt [15] was that for large N, the $GI/M/N$ queue will, for stretches of time when the number of customers is low, behave as if it were an $GI/M/\infty$ queue. Our main result in Section 2.4 shows that the same holds true for the $G/GI/N$ queue as well. Our first step towards showing this is to write down the system equations for the $G/GI/N$ queue in a similar way to those for the $G/GI/\infty$ queue. For the reader's convenience, we will closely adhere to the notation used in [28] as many of the arguments we use here cite results from that paper. All random variables and processes considered

henceforth are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Also, when writing Lesbegue-Stieltjes integrals, we denote by $\int_a^b$ the integral over the closed interval $[a, b]$ and by $\int_{a+}^b$ the integral over the half-open interval $(a, b]$.

Customers arrive to the system according to the arrival counting process $A = \{A(t), t \geq 0\}$ and are served on a first come first served (FCFS) basis. The arrival time of the i^{th} customer is defined to be the quantity

$$\tau_i \quad = \quad \inf\{t \geq 0 : A(t) \geq i\}, \quad i \geq 1.$$

Setting $\tau_0 = 0$, we also define

$$\xi_i \quad = \quad \tau_i - \tau_{i-1}, \quad i \geq 1, \tag{8}$$

to be the interarrival times between the $(i-1)^{st}$ and i^{th} customers to arrive to the system.

The i^{th} customer to enter service after time zero is assigned the service time η_i. We assume that $\{\eta_i, i \geq 1\}$ is an i.i.d. sequence of mean 1, random variables with common distribution F whose tail distribution we denote by $G = 1 - F$. Note that there are no assumptions imposed on the service time distribution, other than it have a finite first moment. Initially, there will also be Q_0 customers in the system and $\min(Q_0, N)$ customers in service at time zero. We let $\tilde{\eta}_i$ be the residual service time of the i^{th} such customer and we assume that $\{\tilde{\eta}_i, i \geq 1\}$ forms an i.i.d. sequence of random variables with common distribution H.

For each $i \geq 1$, let w_i denote the waiting time of the i^{th} customer to arrive to the system and let $\tilde{w}_i$ be the waiting time of the $(N+i)^{th}$ initial customer in the N^{th} system if such a customer exists. We begin our indexing by $N+1$ since the first N initial customers in the system will not have to wait. In this notation as well as the notation of the previous subsection, the total number of customers in the system at time t is given by

$$\begin{aligned} Q(t) \quad = \quad & \sum_{i=1}^{\min(Q_0,N)} 1\{\tilde{\eta}_i > t\} + \sum_{i=1}^{(Q_0-N)^+} 1\{\tilde{w}_i + \eta_i > t\} \\ & + \sum_{i=1}^{A(t)} 1\{\tau_i + w_i + \eta_{(Q_0-N)^+ + i} > t\}. \end{aligned} \tag{9}$$

We henceforth refer to the process $Q = \{Q(t), t \geq 0\}$ as the queue length process. It should be noted that Q does not only count those customers in the queue waiting to be served but that indeed it counts the total number of customers in the system. The number of customers waiting to be served may however be recovered from Q and is given by $(Q - N)^+$. Also note that $Q_0 \neq Q(0)$ in general since it is possible for customers to arrive to the system at time zero. One may think of Q_0 as being equal to $Q(0-)$.

By centering each of the indicator functions in the last two summations on the righthand side of (9) by their means conditional on their arrival times and waiting times, we obtain

$$Q(t) = W(t) + M_2(t) + \sum_{i=1}^{(Q_0-N)^+} G(t - \tilde{w}_i) + \sum_{i=1}^{A(t)} G(t - \tau_i - w_i), \tag{10}$$

where

$$W(t) = \sum_{i=1}^{\min(Q_0,N)} 1\{\tilde{\eta}_i > t\} \tag{11}$$

and

$$\begin{aligned} M_2(t) = & \sum_{i=1}^{(Q_0-N)^+} (1\{\tilde{w}_i + \eta_i > t\} - G(t - \tilde{w}_i)) \\ & + \sum_{i=1}^{A(t)} (1\{\tau_i + w_i + \eta_{(Q_0-N)^+ + i} > t\} - G(t - \tau_i - w_i)). \end{aligned} \tag{12}$$

We also set $W = \{W(t), t \geq 0\}$ and $M_2 = \{M_2(t), t \geq 0\}$.

Next, adding in and subtracting out the terms

$$A_G(t) = \int_0^t G(t - s) dA(s)$$

and $(Q_0 - N)^+ G(t)$ both to and from the righthand side of (10), we obtain

$$\begin{aligned} Q(t) = & W(t) + M_2(t) + A_G(t) + (Q_0 - N)^+ G(t) \\ & + \sum_{i=1}^{(Q_0-N)^+} (G(t - \tilde{w}_i) - G(t)) \\ & + \sum_{i=1}^{A(t)} (G(t - \tau_i - w_i) - G(t - \tau_i)). \end{aligned} \tag{13}$$

We set $A_G = \{A_G(t), t \geq 0\}$.

We now have the following key proposition.

Proposition 2.1.1. *For each $t \geq 0$,*

$$\sum_{i=1}^{A(t)}(G(t-\tau_i-w_i)-G(t-\tau_i)) =$$
$$\int_0^t (Q(t-s)-N)^+ dF(s) - \sum_{i=1}^{(Q_0-N)^+}(G(t-\tilde{w}_i)-G(t)).$$

Proof. We have

$$\begin{aligned}
&= \sum_{i=1}^{A(t)}(G(t-\tau_i-w_i)-G(t-\tau_i)) \\
&= \sum_{i=1}^{A(t)} \int_{(t-(\tau_i+w_i))^+}^{t-\tau_i} dF(s) \\
&= \sum_{i=1}^{A(t)} \int_0^{\infty} 1\{t-(\tau_i+w_i) < s \leq t-\tau_i\} dF(s) \\
&= \sum_{i=1}^{A(t)} \int_0^{\infty} 1\{\tau_i \leq t-s < \tau_i+w_i\} dF(s) \\
&= \int_0^{\infty} \sum_{i=1}^{A(t)} 1\{\tau_i \leq t-s < \tau_i+w_i\} dF(s) \\
&= \int_0^t \left((Q(t-s)-N)^+ - \sum_{i=1}^{(Q_0-N)^+} 1\{\tilde{w}_i > t-s\} \right) dF(s) \\
&= \int_0^t (Q(t-s)-N)^+ dF(s) - \int_0^t \sum_{i=1}^{(Q_0-N)^+} 1\{\tilde{w}_i > t-s\} dF(s).
\end{aligned}$$

A reverse argument can now also be used to show that

$$\int_0^t \sum_{i=1}^{(Q_0-N)^+} 1\{\tilde{w}_i > t-s\} dF(s) = \sum_{i=1}^{(Q_0-N)^+}(G(t-\tilde{w}_i)-G(t)).$$

This completes the proof. □

Proposition 2.1.1 now allows us to rewrite equation (13) for the queue length at time t as

$$\begin{aligned}
Q(t) &= W(t) + M_2(t) + A_G(t) + (Q_0-N)^+ G(t) \\
&\quad + \int_0^t (Q(t-s)-N)^+ dF(s).
\end{aligned} \tag{14}$$

Equation (14) is the starting point for our analysis in Sections 2.3 and 2.4. In the next section, we develop a regulator map result for the queue length process in (14).

2.1.2 The Halfin-Whitt Heavy Traffic Asymptotic Regime

In this subsection, the details of our heavy traffic formulation are given. The underlying assumption is that we are considering a sequence of $G/GI/N$ queueing systems which are indexed by the number of servers N. In what follows, stochastic processes are assumed to be measurable maps from $(\Omega, \mathcal{F})$ to $(D[0,\infty), \mathcal{D})$, where $\mathcal{D}$ stands for the σ-algebra generated by the Skorohod J_1 topology. We will use d_0 to denote the Skorohod metric. The notation $D^k, k \geq 1$, will be used to denote the product space of D defined by $D^k = D \times \ldots \times D$. In general, when speaking of product spaces, we will assume that they are endowed with the product topology which is induced by the maximum metric d. The interested reader is referred to Section 11.4 of [54] for futher details. The notation $\Rightarrow$ will be used to denote convergence in distribution and the identity process $e = \{t, t \geq 0\}$ will be denoted by e.

Let $A^N = \{A^N(t), t \geq 0\}$ be the arrival process to our N^{th} system. Defining the family of fluid scaled arrival processes $\left\{\bar{A}^N,\ N \geq 1\right\}$, where

$$\bar{A}^N(t) = \frac{A^N(t)}{N},$$

and $\bar{A}^N = \{\bar{A}^N(t), t \geq 0\}$, we assume that

$$\bar{A}^N \Rightarrow e \ \text{ as } \ N \to \infty. \tag{15}$$

It will also be assumed that the fluctuations of our arrival processes around their means obeys some form of a functional central limit theorem. That is, there exists a sequence of constants $\rho^N,\ N \geq 1$, such that

$$\tilde{A}^N \Rightarrow \xi \ \text{ as } N \to \infty, \tag{16}$$

where

$$\tilde{A}^N(t) = \sqrt{N}(\bar{A}^N(t) - \rho^N t) \tag{17}$$

and $\tilde{A}^N = \{\tilde{A}^N(t), t \geq 0\}$. Furthermore, we will also assume that the limiting process ξ in (16) has almost surely continuous sample paths.

Note that assumption (16) is flexible from a modeling point of view. In heavy traffic theory, one might often assumes that $A^N(e) = A(Ne)$, where A is a renewal process, in

which case, by Donsker's Theorem, the process ξ in (16) turns out to be a Brownian motion. The interpretation under such an assumption is that customers are emanating from a single source, albeit at a rapid rate. However, in many applications, with telephone call centers being just one such example, it is perhaps more natural to assume that customers are emanating from many sources. This then leads to the assumption that A^N is a superposition of many i.i.d. renewal arrival processes, that is, $A^N = \sum_{i=1}^{N} A_i$. Under such an assumption, the process ξ turns out to be a centered Gaussian process whose covariance structure is inherited from that of each of the individual $A_i's$. The interested reader may consult Section 7.2 of Whitt [54] for more details on this remark.

The service time sequences do not change as we index through N. In other words, the sequence $\{\eta_i, i \geq 1\}$ will always be used to assign service times to customers entering service after time zero and $\{\tilde{\eta}_i, i \geq 1\}$ will also be used to denote the initial residual service time sequence. We will let w_i^N be the waiting time of the i^{th} customer to arrive to the N^{th} system after time zero and denote by $\tilde{w}_i^N$ the waiting time of the $(N+i)^{th}$ initial customer, if any.

It now remains to define the Halfin-Whitt regime. First note that in the case that $A^N(t) = A(N\rho^N t), t \geq 0$, where $A = \{A(t), t \geq 0\}$ is a renewal process with mean 1 interarrival times, we have that $N\rho^N$ is the arrival rate to the N^{th} system. In this case, as noted previously, the limiting process ξ in (16) is a Brownian motion and the traffic intensity of the N^{th} system is ρ^N since there are N servers in the system and each has rate 1 processing capacity. For more general arrival processes, however, the quantity $N\rho^N$ may fail to be the arrival rate to the N^{th} system. Nevertheless, the Halfin-Whitt regime is achieved by assuming that ρ^N converges to one as N grows to infinity. Specifically, we assume that

$$\sqrt{N}(1 - \rho^N) \to \beta \quad \text{as} \quad N \to \infty, \tag{18}$$

where $-\infty < \beta < \infty$.

2.2 A Regulator Map Result

In this section, a regulator map result is provided which will be relied upon in the proof of our main result. In particular, it will provide a convenient representation for the queue length process.

Let B be a cumulative distribution function on $\mathbb{R}$. For each $x \in D[0,\infty)$, we would like to find and characterize solutions $z \in D[0,\infty)$ to equations of the form

$$z(t) = x(t) + \int_0^t z^+(t-s)dB(s), \quad \text{for} \quad t \geq 0. \tag{19}$$

We therefore define the mapping $\varphi_B : D[0,\infty) \mapsto D[0,\infty)$ to be such that $\varphi_B(x)$ is a solution to (19) for each $x \in D[0,\infty)$. The following proposition now shows that φ_B is uniquely defined and provides some regularity results for φ_B as well. Its proof may be found in the appendix.

Proposition 2.2.1. *For each $x \in D[0,\infty)$, there exists a unique solution $\varphi_B(x)$ to (19). Moreover, the function $\varphi_B : D[0,\infty) \mapsto D[0,\infty)$ is Lipschitz continuous in the topology of uniform convergence over bounded intervals and measurable with respect to the Skorohod J_1 topology.*

2.3 Fluid Limit Results

In this section, we study the fluid scaled queue length process. This will be done as follows. We first represent the queue length process in term of the regulator mapping φ_F. We then provide several propositions which will be helpful in establishing our main result of the section, Theorem 2.3.4, which details the asymptotic behavior of the fluid scaled queue length process. We conclude this section by showing in Theorem 2.3.6 that the fluid limit for the queue length process converges to a unique equilibrium state as time increases to infinity.

Let $Q^N = \{Q^N(t), t \geq 0\}$ be the queue length process in the N^{th} system. First note that by setting

$$I^N(t) \quad = \quad (N - (Q_0^N - N)^-)\bar{H}(t) + (Q_0^N - N)^+ G(t)$$

and

$$W_0^N(t) = \sum_{i=1}^{\min(Q_0^N,N)} (1\{\tilde{\eta}_i \geq t\} - \bar{H}(t)),$$

we have, after subtracting N from both the left and righthand sides of (14), that

$$\begin{aligned} Q^N(t) - N &= I^N(t) + W_0^N(t) + M_2^N(t) + (A_G^N(t) - N) \\ &\quad + \int_0^t (Q^N(t-s) - N)^+ dF(s). \end{aligned} \tag{20}$$

If we now define the fluid scaled quantities,

$$\bar{Q}^N(t) = \frac{Q^N(t) - N}{N}, \tag{21}$$

$$\begin{aligned} \bar{I}^N(t) &= \frac{I^N(t)}{N}, \\ \bar{W}_0^N(t) &= \frac{W_0^N(t)}{N}, \\ \bar{M}_2^N(t) &= \frac{M_2^N(t)}{N} \end{aligned} \tag{22}$$

and

$$\bar{A}_G^N(t) = \frac{A_G^N(t) - N}{N}, \tag{23}$$

it then follows from (20) that

$$\bar{Q}^N(t) = \bar{I}^N(t) + \bar{W}_0^N(t) + \bar{M}_2^N(t) + \bar{A}_G^N(t) + \int_0^t \bar{Q}^{N,+}(t-s) dF(s). \tag{24}$$

Furthermore, since by Proposition 2.2.1, the mapping φ_F is uniquely defined, setting

$$\begin{aligned} \bar{Q}^N &= \{\bar{Q}^N(t), t \geq 0\}, \\ \bar{I}^N &= \{\bar{I}^N(t), t \geq 0\}, \\ \bar{W}_0^N &= \{\bar{W}_0^N(t), t \geq 0\}, \\ \bar{M}_2^N &= \{\bar{M}_2^N(t), t \geq 0\} \end{aligned}$$

and

$$\bar{A}_G^N(t) = \{\bar{A}_G^N(t), t \geq 0\},$$

we have from (24) that

$$\bar{Q}^N \quad = \quad \varphi_F(\bar{I}^N + \bar{W}_0^N + \bar{M}_2^N + \bar{A}_G^N). \tag{25}$$

The representation (25) above will turn out to be very useful when proving the main result of this section.

For the remainder of this chapter, we assume that the initial fluid scaled number of customers in the system converges in distribution as N tends to ∞. In other words, we assume that

$$\bar{Q}_0^N = \frac{Q_0^N - N}{N} \Rightarrow \bar{Q}_0 \quad \text{as} \quad N \to \infty. \tag{26}$$

We may now begin to state some preliminary results in preparation for the statement of the main result of the section, Theorem 2.3.4. Our first result shows that $\bar{W}_0^N$ converges to zero as N goes to ∞.

Proposition 2.3.1. *$\bar{W}_0^N \Rightarrow 0$ as $N \to \infty$.*

Proof. First note that

$$\bar{W}_0^N(t) = N^{-1} \sum_{i=1}^{N \min(N^{-1}Q_0^N, 1)} (1\{\tilde{\eta}_i > t\} - \bar{H}(t)).$$

Thus, for each $T > 0$ and $\delta > 0$, we have

$$\begin{aligned} & P\left(\sup_{0 \le t \le T} \left| N^{-1} \sum_{i=1}^{N \min(N^{-1}Q_0^N, 1)} (1\{\tilde{\eta}_i > t\} - \bar{H}(t)) \right| > \delta \right) \\ \le \quad & P\left(\sup_{0 \le x \le 1} \sup_{0 \le t \le T} \left| N^{-1} \sum_{i=1}^{\lfloor xN \rfloor} (1\{\tilde{\eta}_i > t\} - \bar{H}(t)) \right| > \delta \right). \end{aligned}$$

However, by Lemma 3.1 in [28],

$$P\left(\sup_{0 \le x \le 1} \sup_{0 \le t \le T} \left| N^{-1} \sum_{i=1}^{\lfloor xN \rfloor} (1\{\tilde{\eta}_i > t\} - \bar{H}(t)) \right| > \delta \right) \to 0 \quad \text{as} \quad N \to \infty,$$

which completes the proof. □

We next show that $\bar{M}_2^N$ converges in distribution to zero. The proof of this result may be found in the appendix.

Proposition 2.3.2. *$\bar{M}_2^N \Rightarrow 0$ as $N \to \infty$.*

Proof. See appendix. □

Now define

$$F_e(t) = \int_0^t G(t-s)ds, \quad t \geq 0, \tag{27}$$

to be the equilibrium distribution associated with the service time distribution F and set $\bar{F}_e = 1 - F_e$ to be the tail distribution of F_e. We then have the following result.

Proposition 2.3.3. *$\bar{A}_G^N \Rightarrow -\bar{F}_e$ as $N \to \infty$.*

Proof. By the definition of $\bar{A}_G^N$ in (23), assumption (15) and as in the proof of Theorem 3 of [28],

$$\begin{aligned} \bar{A}_G^N &= \int_0^{\cdot} G(\cdot - s)d\bar{A}^N(s) - 1 \\ &\Rightarrow \int_0^{\cdot} G(\cdot - s)ds - 1 \quad \text{as} \quad N \to \infty \\ &= -\bar{F}_e, \end{aligned}$$

which completes the proof. □

The following is now the main result of this section. It provides a deterministic first order approximation to the queue length process. Later, in Section 2.4, we use this result to center the queue length process and obtain a second order approximation.

Theorem 2.3.4. *If $\bar{Q}_0^N \Rightarrow \bar{Q}_0$ as $N \to \infty$, then $\bar{Q}^N \Rightarrow \bar{Q}$ as $N \to \infty$, where $\bar{Q}$ is the unique strong solution to*

$$\bar{Q}(t) = (1 - \bar{Q}_0^-)\bar{H}(t) + \bar{Q}_0^+ G(t) - \bar{F}_e(t) + \int_0^t \bar{Q}^+(t-s)dF(s), \quad \textit{for} \quad t \geq 0. \tag{28}$$

Proof. Setting

$$\bar{M}_3^N = \bar{W}_0^N + \bar{M}_2^N + \bar{A}_G^N,$$

it follows by Propositions 2.3.1, 2.3.2 and 2.3.3 that

$$\bar{M}_3^N \Rightarrow -\bar{F}_e \quad \text{as} \quad N \to \infty.$$

Since, by assumption, $\bar{Q}_0^N \Rightarrow \bar{Q}_0$ as $N \to \infty$, it now follows by Theorem 11.4.5 in [54] that

$$(\bar{M}_3^N, \bar{Q}_0^N) \Rightarrow (-\bar{F}_e, \bar{Q}_0) \text{ in } D \times \mathbb{R} \text{ as } N \to \infty.$$

By Theorem 11.4.1 in [54], the space $D \times \mathbb{R}$ is separable under the product topology and thus, by the Skorohod representation theorem [54], there exists some alternate probability space, $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$, on which are defined a sequence of processes

$$\{(\hat{M}_3^N, \hat{Q}_0^N), N \geq 1\} \tag{29}$$

such that

$$(\hat{M}_3^N, \hat{Q}_0^N) \stackrel{d}{=} (\bar{M}_3^N, \bar{Q}_0^N) \text{ for } N \geq 1, \tag{30}$$

and also processes

$$(-\bar{F}_e, \hat{Q}_0) \stackrel{d}{=} (-\bar{F}_e, \bar{Q}_0), \tag{31}$$

where

$$(\hat{M}_3^N, \hat{Q}_0^N) \to (-\bar{F}_e, \hat{Q}_0) \ \hat{\mathbb{P}} \text{ a.s. as } N \to \infty. \tag{32}$$

Furthermore, as the process $-\bar{F}_e$ on the righthand side of (32) is, by (27), continuous, it follows that the convergence in (32) can also be strengthened to uniform convergence on compact sets (u.o.c.).

Now set

$$\hat{I}^N = (1 - \hat{Q}_0^{N,-})\bar{H} + \hat{Q}_0^{N,+} G$$

and note that by (30), we have

$$(\hat{M}_3^N, \hat{I}^N) \stackrel{d}{=} (\bar{M}_3^N, \bar{I}^N) \text{ for } N \geq 1. \tag{33}$$

Furthermore, letting

$$\hat{I} = (1 - \hat{Q}_0^-)\bar{H} + \hat{Q}_0^+ G,$$

we have for each $T \geq 0$, by (32),

$$\begin{aligned}
\sup_{0\leq t\leq T} |\hat{I}^N(t) - \hat{I}(t)| &= \sup_{0\leq t\leq T} |(\hat{Q}_0^- - \hat{Q}_0^{N,-})\bar{H}(t) + (\hat{Q}_0^{N,+} - \hat{Q}_0^+)G(t)| \\
&\leq |\hat{Q}_0^N - \hat{Q}_0| \sup_{0\leq t\leq T} (\bar{H}(t) + G(t)) \\
&\leq 2|\hat{Q}_0^N - \hat{Q}_0| \\
&\to 0 \ \ \hat{\mathbb{P}} \text{ a.s. as } N \to \infty. \qquad (34)
\end{aligned}$$

Now let

$$\hat{Q}^N = \varphi_F(\hat{I}^N + \hat{M}_3^N)$$

and note that by the representation (25), (29), (33) and the measurability of φ_F from Proposition 2.2.1, it follows that

$$\hat{Q}^N \stackrel{d}{=} \tilde{Q}^N \ \text{ for } \ N \geq 1.$$

Furthermore, it follows from (32), (34) and the continuity of φ_F with respect to the topology of uniform convergence over compact sets from Proposition 2, that

$$\hat{Q}^N = \varphi_F(\hat{I}^N + \hat{M}_3^N) \to \varphi_F((1 + \hat{Q}_0^-)\bar{H} + \hat{Q}_0^+ G - \bar{F}_e) \ \text{ as } \ N \to \infty,$$

u.o.c. $\hat{\mathbb{P}}$ a.s. This completes the proof. □

By Theorem 2.3.4, the following result is now trivial. It is revisited again in the Section that follows.

Corollary 2.3.5. *Suppose that $\bar{Q}_0^N \Rightarrow 0$ as $N \to \infty$ and that the initial residual service time distribution is given by F_e. Then, $\bar{Q}^N \Rightarrow 0$ as $N \to \infty$.*

Proof. Substituting the initial conditions $\bar{Q}_0 = 0$ and $\bar{H} = \bar{F}_e$ into equation (28) for the limiting fluid scaled queue length process $\bar{Q}$, we obtain the equation

$$\bar{Q}(t) = 0 + \int_0^t \bar{Q}^+(t-s)dF(s), \ \text{ for } \ t \geq 0,$$

which has the unqiue solution $\bar{Q} = 0$. This completes the proof. □

We next turn to analyzing the fluid limit in (28) in more detail. The following result shows that regardless on the initial configuration of the system, it eventually reaches a steady state on the fluid scale.

Theorem 2.3.6. *Suppose that both the initial service time distribution H and the equilibrium distribution F_e posses a finite mean and that the service time distribution F is non-lattice. Let $\bar{Q}$ be the fluid limit obtained from Theorem 2.3.4 above. Then, there exists a constant $c^\star \geq 0$, which in general depends on the initial conditions $\bar{Q}_0$ and H of the system, such that*

$$\lim_{t\to\infty} \bar{Q}(t) = c^\star.$$

Proof. First note by (28), that for $t \geq 0$,

$$\begin{aligned} \bar{Q}(t) &= (1-\bar{Q}_0^-)\bar{H}(t) + \bar{Q}_0^+ G(t) - \bar{F}_e(t) + \int_0^t \bar{Q}^+(t-s)dF(s) \qquad (35) \\ &\geq (1-\bar{Q}_0^-)\bar{H}(t) + \bar{Q}_0^+ G(t) - \bar{F}_e(t) \\ &\to 0 \text{ as } t\to\infty \end{aligned}$$

and so it follows that

$$\bar{Q}^-(t) \to 0 \text{ as } t\to\infty.$$

It remains to find a limit for $\bar{Q}^+$. Since $\bar{Q} = \bar{Q}^+ - \bar{Q}^-$, first note by (35) that

$$\bar{Q}^+(t) = \bar{X}(t) + \int_0^t \bar{Q}^+(t-s)dF(s)$$

where

$$\bar{X}(t) = -\bar{Q}^-(t) + (1-\bar{Q}_0^-)\bar{H}(t) + \bar{Q}_0^+ G(t) - \bar{F}_e(t). \qquad (36)$$

Thus, if we may show that $\bar{X}$ is directly Riemann integrable, then it will follow by the key renewal theorem that

$$\lim_{t\to\infty} \bar{Q}^+(t) = \int_0^\infty \bar{X}(s)ds,$$

which will complete the proof.

We will now show that $\bar{X}$ is directly Riemann integrable. Note that since by assumption H and F both posses finite means, it follows that

$$(1-\bar{Q}_0^-)\bar{H} + \bar{Q}_0^+ G$$

is directly Riemann integrable, being the sum of two nonincreasing Riemann integrable functions. Next, since F_e is assumed to have a finite mean, it also follows that $\bar{F}_e$ is directly Reimann integrable from which follows the direct Riemann integrability of

$$(1 - \bar{Q}_0^- \bar{H} + \bar{Q}_0^+ G - \bar{F}_e.$$

Thus, in order to complete the proof, it is sufficient to show that $\bar{Q}^-$ a directly Riemann integrable function. However, this follows since by (35),

$$0 \leq \bar{Q}^- \leq \bar{F}_e$$

and $\bar{F}_e$ is a directly Riemann integrable function. The proof is now complete. □

2.4 Diffusion Limit Results

In this section, we obtain limiting results for the diffusion scaled queue length process. This is accomplished by first writing system equation (20) for the queue length process in terms of the regulator map φ_F defined in Section 2.2. We then provide several useful propositions and lemmas in preparation for the statement of our main result, Theorem 2.4.4, which provides a limiting approximation to the diffusion scaled queue length process. Corollary 2.4.5 next provides an alternative representation of the limiting process in Theorem 2.4.4. We conclude by showing how the alternative representation of Corollary 2.4.5 reduces to the diffusion obtained by Halfin and Whitt [18] in the case of exponentially distributed service times and renewal arrivals.

For the remainder of this section, we will make the simplifying assumption that the residual service time distribution of those customers being served at time zero is equal to F_e, the equilibrium distribution associated with F and that fluid scaled initial number of customers in the system converges to zero, i.e.

$$\bar{Q}_0^N \Rightarrow 0 \quad \text{as} \quad N \to \infty.$$

We will also assume that on a diffusion scale, the initial number of customers converges. That is,

$$\tilde{Q}^N(0) = \frac{Q_0^N - N}{\sqrt{N}} \Rightarrow \tilde{Q}(0) \quad \text{as} \quad N \to \infty. \tag{37}$$

Recall now that equation (20) in Section 2.3 shows that the queue length at time t may be written as

$$\begin{aligned} Q^N(t) &= I^N(t) + W_0^N(t) + M_2^N(t) + A_G^N(t) \\ &\quad + \int_0^t (Q^N(t-s) - N)^+ dF(s). \end{aligned}$$

Centering the above equation by $N + N\bar{Q}(t) = N$ and performing some algebraic manipulations, one then obtains that

$$\begin{aligned} Q^N(t) - N &= M_Q^N(t) + H^N(t) + W_0^N(t) + M_2^N(t) + M_1^N(t) \\ &\quad + \int_0^t (Q^N(t-s) - N)^+ dF(s), \end{aligned} \tag{38}$$

where

$$M_Q^N(t) = (Q_0^N - N)^+ (G(t) - \bar{F}_e(t)),$$

$$H^N(t) = (Q_0^N - N)\bar{F}_e(t)$$

and

$$M_1^N(t) = \int_0^t G(t-s) d(A^N(s) - Ns). \tag{39}$$

Let $M_Q^N = \{M_Q^N(t), t \geq 0\}$, $H^N = \{H^N(t), t \geq 0\}$ and $M_1^N = \{M_1^N(t), t \geq 0\}$.

If we now define the diffusion scaled quantities,

$$\tilde{Q}^N(t) = \frac{Q^N(t) - N}{\sqrt{N}}, \tag{40}$$

$$\begin{aligned} \tilde{M}_Q^N(t) &= \frac{M_Q^N(t)}{\sqrt{N}}, \\ \tilde{H}^N(t) &= \frac{H^N(t)}{\sqrt{N}}, \\ \tilde{W}_0^N(t) &= \frac{W_0^N(t)}{\sqrt{N}}, \\ \tilde{M}_2^N(t) &= \frac{M_2^N(t)}{\sqrt{N}} \end{aligned} \tag{41}$$

and

$$\tilde{M}_1^N(t) = \frac{M_1^N(t)}{\sqrt{N}},$$

it then follows from (38), that

$$\tilde{Q}^N(t) \;=\; \tilde{M}_Q^N(t) + \tilde{Q}_I^N(t) + \int_0^t \tilde{Q}^{N,+}(t-s)dF(s), \tag{42}$$

where

$$\tilde{Q}_I^N(t) = \tilde{H}^N(t) + \tilde{W}_0^N(t) + \tilde{M}_1^N(t) + \tilde{M}_2^N(t).$$

Letting

$$\begin{aligned}
\tilde{Q}^N &= \{\tilde{Q}^N(t), t \geq 0\},\\
\tilde{M}_Q^N &= \{\tilde{M}_Q^N(t), t \geq 0\},\\
\tilde{H}^N &= \{\tilde{H}^N(t), t \geq 0\},\\
\tilde{W}_0^N &= \{\tilde{W}_0^N(t), t \geq 0\},\\
\tilde{M}_2^N &= \{\tilde{M}_2^N(t), t \geq 0\},\\
\tilde{M}_1^N(t) &= \{\tilde{M}_1^N(t), t \geq 0\}
\end{aligned}$$

and

$$\tilde{Q}_I^N = \tilde{H}^N + \tilde{W}_0^N + \tilde{M}_1^N + \tilde{M}_2^N,$$

we then have, since the mapping φ_F is by Proposition 2.2.1 uniquely defined, that (42) may also be written as

$$\tilde{Q}^N \;=\; \varphi_F(\tilde{M}_Q^N + \tilde{Q}_I^N). \tag{43}$$

The representation (43) will be useful in the proof our main result. However, before stating this result, we first provide several preliminary propositions and lemmas which are interesting in their own right.

Let $\hat{A}^N(t)$ be the number of customers who have entered service by time t, excluding those customers who were intially in service at time zero, and set $\hat{A}^N = \{\hat{A}^N(t), t \geq 0\}$. We then have the following.

Lemma 2.4.1. *$N^{-1}\hat{A}^N \Rightarrow e$ as $N \to \infty$.*

Proof. First note the relationship

$$\hat{A}^N(t) = (Q^N(0) - N)^+ + A^N(t) - (Q^N(t) - N)^+. \tag{44}$$

It then follows by assumption (15), Corollary 2.3.5, assumption (37) and the continuous mapping theorem that

$$N^{-1}\hat{A}^N = \bar{Q}^{N,+}(0) + \bar{A}^N - \bar{Q}^{N,+} \Rightarrow e \quad \text{as} \quad N \to \infty,$$

which completes the proof. □

We next provide a result which will be crucial in the proof of our main result. Its proof may be found in the appendix.

Proposition 2.4.2. *Let*

$$\hat{M}_2^N(t) = N^{-1/2} \sum_{i=1}^{Nt} (1\{N^{-1}i + \eta_i \geq t\} - G(t - N^{-1}i)), \quad t \geq 0,$$

and set $\hat{M}_2^N = \{\hat{M}_2^N(t), t \geq 0\}$. *Then,*

$$(\tilde{M}_2^N, \hat{M}_2^N) \Rightarrow (\tilde{M}_2, \tilde{M}_2) \quad as \quad N \to \infty,$$

where $\tilde{M}_2$ *is a centered Gaussian process with covariance structure*

$$E[\tilde{M}_2(t)\tilde{M}_2(t+\delta)] = \int_0^t G(t+\delta-u)F(t-u)du \quad for \quad t, \delta \geq 0.$$

Proof. See appendix. □

We now prove a joint convergence result on the diffusion scaled processes defined at the beginning of this section. Let

$$\tilde{M}_1(t) = \int_0^t G(t-s)d(\xi(s) - \beta s), \quad t \geq 0, \tag{45}$$

and set $\tilde{M}_1 = \{\tilde{M}_1(t), t \geq 0\}$. The process ξ appearing in the definition of $\tilde{M}_1$ above is the limiting process appearing in (16) and β arises from the QED condition (18). Note also that the integral above may be interpreted as integration by parts.

Next, let $\tilde{W}_0 = \{\tilde{W}_0(t), t \geq 0\}$ be a Brownian bridge. In other words, $\tilde{W}_0$ is the unique continuous, centered Gaussian process on $[0,1]$ with covariance function

$$E[\tilde{W}_0(s)\tilde{W}_0(t)] \quad = \quad (s \wedge t) - st, \quad 0 \leq s \leq t \leq 1.$$

Moreover, set $\tilde{W}_0(F_e) = \{\tilde{W}_0(F_e(t)), t \geq 0\}$, where F_e is the equilibrium distribution associated with F as defined in (27). One may view $\tilde{W}_0(F_e)$ as a time changed Brownian bridge. We then have the following result.

Proposition 2.4.3.

$$(\tilde{Q}^N(0), \tilde{W}_0^N, \tilde{M}_1^N, \tilde{M}_2^N) \Rightarrow (\tilde{Q}(0), \tilde{W}_0(F_e), \tilde{M}_1, \tilde{M}_2) \quad in \quad \mathbb{R} \times D^3$$

as $N \to \infty$, where each of the limiting processes appearing on the righthand side above are independent of one another.

Proof. We first show convergence of the marginals. The convergence of $\tilde{Q}^N(0)$ to $\tilde{Q}(0)$ is clear by assumption (37).

Let

$$\hat{W}^N(t) = N^{-1/2} \sum_{i=1}^{N} (1\{\tilde{\eta}_i > t\} - \bar{F}_e(t))$$

and set $\hat{W}^N = \{\hat{W}^N(t), t \geq 0\}$. The convergence

$$(\hat{W}_0^N, \tilde{W}_0^N) \quad \Rightarrow \quad (\tilde{W}_0(F_e), \tilde{W}_0(F_e)) \quad \text{as} \quad N \to \infty \tag{46}$$

follows by the representation

$$\tilde{W}^N(t) = N^{-1/2} \sum_{i=1}^{N(\min(N^{-1}Q_0^N, 1)} (1\{\tilde{\eta}_i > t\} - \bar{F}_e(t)),$$

the random time change theorem and Lemma 3.1 of [28], since, by the continuous mapping theorem and assumption (37),

$$(\min(N^{-1}Q_0^N, 1) \Rightarrow 1 \quad \text{as} \quad N \to \infty.$$

Finally, the convergence of $\tilde{M}_1^N$ to $\tilde{M}_1$ follows by Theorem 3 of [28] and the convergence of $\tilde{M}_2^N$ to $\tilde{M}_2$ is immediate by Proposition 2.4.2.

It remains to show the joint convergence as stated in the proposition. The convergence

$$(\tilde{Q}^N(0), \hat{W}_0^N, \tilde{M}_1^N, \hat{M}_2^N) \Rightarrow (\tilde{Q}(0), \tilde{W}_0(F_e), \tilde{M}_1, \tilde{M}_2) \text{ in } \mathbb{R} \times D^3$$

as $N \to \infty$ follows by Theorem 11.4.4 in [54] since each of the component processes appearing in the prelimit above are independent of one another and further, they converge to their desired limits as shown in the previous paragraph. Next, note that

$$d((\tilde{Q}^N(0), \tilde{W}_0^N, \tilde{M}_1^N, \tilde{M}_2^N), (\tilde{Q}^N(0), \hat{W}_0^N, \tilde{M}_1^N, \hat{M}_2^N)) \leq d_0(\tilde{W}_0^N, \hat{W}_0^N) + d_0(\tilde{M}_2^N, \hat{M}_2^N)$$

and thus, if we can show that

$$d_0(\tilde{W}_0^N, \hat{W}_0^N) + d_0(\tilde{M}_2^N, \hat{M}_2^N) \Rightarrow 0 \text{ as } N \to \infty, \tag{47}$$

then by Theorem 11.4.7 in [54] the proof will be complete. However, (47) follows by (46), Proposition 2.4.2 and Theorem 11.4.8 in [54]. The proof is now complete. □

We are now ready to state the main result of this section. Let

$$\tilde{H} = \tilde{Q}(0)\bar{F}_e \quad \text{and} \quad \tilde{M}_Q = \tilde{Q}^+(0)(G - \bar{F}_e). \tag{48}$$

Furthermore, set

$$\tilde{Q}_I = \tilde{H} + \tilde{W}_0(F_e) + \tilde{M}_1 + \tilde{M}_2. \tag{49}$$

Note that by Theorem 3 of [28], $\tilde{Q}_I$ is the limiting queue length process of a sequence of $G/GI/\infty$ queues with identical arrival processes and service time sequence as our original sequence of $G/GI/N$ queues and with Q_0^N customers in service at time zero with residual service time distribution F_e.

The following is now our second main result.

Theorem 2.4.4. *If the residual service time distribution $H = F_e$ and $\tilde{Q}^N(0) \Rightarrow \tilde{Q}(0)$ as $N \to \infty$, then $\tilde{Q}^N \Rightarrow \varphi_F(\tilde{M}_Q + \tilde{Q}_I)$ as $N \to \infty$.*

Proof. By Proposition 2.4.3, we have that

$$(\tilde{Q}^N(0), \tilde{W}_0^N, \tilde{M}_1^N, \tilde{M}_2^N) \Rightarrow (\tilde{Q}(0), \tilde{W}_0(F_e), \tilde{M}_1, \tilde{M}_2) \text{ in } \mathbb{R} \times D^3$$

as $N \to \infty$, where each of the limiting processes appearing on the righthand side above are independent of one another. Since $(\mathbb{R}, \mathbb{R})$ and $(D, \mathcal{D})$ are both separable spaces, it follows by Theorem 11.4.1 in [54] that $\mathbb{R} \times D^3$ is separable under the product topology. Thus, by the Skorohod Representation Theorem [54], there exists some alternate probability space, $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$, on which are defined a sequence of processes

$$\{(\hat{Q}_0^N, \hat{W}_0^N, \hat{M}_1^N, \hat{M}_2^N),\ N \geq 1\}$$

where

$$(\hat{Q}_0^N, \hat{W}_0^N, \hat{M}_1^N, \hat{M}_2^N) \stackrel{d}{=} (\tilde{Q}^N(0), \tilde{W}_0^N, \tilde{M}_1^N, \tilde{M}_2^N) \ \text{ for } \ N \geq 1, \tag{50}$$

and also processes

$$(\hat{Q}_0, \hat{W}_0(F_e), \hat{M}_1, \hat{M}_2) \stackrel{d}{=} (\tilde{Q}(0), \tilde{W}_0(F_e), \tilde{M}_1, \tilde{M}_2), \tag{51}$$

such that

$$(\hat{Q}_0^N, \hat{W}_0^N, \hat{M}_1^N, \hat{M}_2^N) \to (\hat{Q}_0, \hat{W}_0(F_e), \hat{M}_1, \hat{M}_2) \ \text{ as } \ N \to \infty, \tag{52}$$

$\hat{\mathbb{P}}$ a.s., where the convergence occurs in the Skorohod J_1-metric. Furthermore, since each of the processes appearing on the righthand side of (52) is continuous, we may assume that above convergence also occurs uniformly on compact sets (u.o.c.)

Now set

$$\hat{M}_Q^N = \hat{Q}^{N,+}(0)(\bar{F} - \bar{F}_e)$$

and

$$\hat{M}_Q = \hat{Q}^{+}(0)(\bar{F} - \bar{F}_e).$$

It is then clear that

$$\begin{aligned} \sup_{0 \leq t \leq T} |\hat{M}_Q^N(t) - \hat{M}_Q(t)| &\leq |\hat{Q}_0^N - \hat{Q}_0| \sup_{0 \leq t \leq T} |\bar{F}(t) - \bar{F}_e(t)| \\ &\leq 2|\hat{Q}_0^N - \hat{Q}_0|, \end{aligned}$$

and so it follows by (52) that

$$\hat{M}_Q^N \to \hat{M}_Q \quad \text{as} \quad N \to \infty \quad \text{u.o.c.} \quad \hat{\mathbb{P}} \quad \text{a.s.} \tag{53}$$

Next, letting

$$\hat{H}^N = \hat{Q}_0^N \bar{F}_e,$$

a similar augment shows that

$$\hat{H}^N \to \hat{H} \quad \text{as} \quad N \to \infty \quad \text{u.o.c.} \quad \hat{\mathbb{P}} \quad \text{a.s.}, \tag{54}$$

where

$$\hat{H} = \hat{Q}_0 \bar{F}_e.$$

Now letting

$$\hat{Q}_I^N = \hat{H}^N + \hat{W}_0^N + \hat{M}_1^N + \hat{M}_2^N,$$

we have by (52), (53) and (54) that

$$\hat{M}_Q^N + \hat{Q}_I^N \to \hat{M}_Q + \hat{Q}_I \quad \text{as} \quad N \to \infty \quad \text{u.o.c.} \quad \hat{\mathbb{P}} \quad \text{a.s.}, \tag{55}$$

where

$$\hat{Q}_I = \hat{H} + \hat{W}(F_e) + \hat{M}_1 + \hat{M}_2.$$

Furthermore, it follows by (51) that

$$\hat{M}_Q^N + \hat{Q}_I^N \stackrel{d}{=} \tilde{M}_Q^N + \tilde{Q}_I^N \quad \text{for} \quad N \geq 1. \tag{56}$$

Now set

$$\hat{Q}^N = \varphi_F(\hat{M}_Q^N + \hat{Q}_I^N). \tag{57}$$

Since by Proposition 2.2.1, the map φ_F is measurable, it follows by (43), (56) and (57) that

$$\hat{Q}^N \stackrel{d}{=} \tilde{Q}^N \quad \text{for} \quad N \geq 1.$$

Furthermore, by the continuity portion of Proposition 2.2.1 and (55),

$$\hat{Q}^N = \varphi_F(\hat{M}_Q^N + \hat{Q}_I^N) \to \varphi_F(\hat{M}_Q + \hat{Q}_I) \text{ as } N \to \infty \text{ u.o.c. } \hat{\mathbb{P}} \text{ a.s.},$$

which, by (56), completes the proof. □

In words, one may view the result of Theorem 2.4.4 as saying that in the limit the queue length process of the $G/GI/N$ queue is equal to that of that of an infinite server queue plus some positive adjustment. It is important to point out that the adjustment is positive by necessity since the number of customers in the $G/GI/N$ system will, because service times are assigned to customers as they arrive, always be sample pathwise more than in the $G/GI/\infty$ queue. Furthermore, the term $\tilde{Q}^+$ appears in the limit since it is only when there are more customers than servers in the system that the finite server approximation to the infinite server queue will be off.

Nevertheless, it is still not immediately clear that the limiting process of Theorem 2.4.4 is equivalent to that of Theorem 2 of Halfin and Whitt [19] in the case of exponentially distributed service times. The following corollary is helpful in showing that this is indeed the case. The intuition behind the corollary is the following. Rather than approximating the system as an infinite server queue, suppose it was simply assumed that all of the servers were constantly running. Of course, this will be the case when there is a positive number of customers waiting to be served. However, if there are no customers waiting to be served, then there will be some servers idle and this approximation will provide too few customer in the system. Thus, whenever $Q^N < N$ it will be necessary to add back some customers to the approximation and the question then becomes how much. In a forthcoming paper [41], the arguments behind this intuition will be be rigorously justified. However, for the present time, we have the following.

Let $M = \{M(t), t \geq 0\}$ be the renewal function associated with the pure renewal process with interarrival distribution given by the service time distribution F. Also, set

$$\zeta = \tilde{M}_Q + \tilde{Q}_I. \tag{58}$$

Corollary 2.4.5. *$\tilde{Q}^N \Rightarrow \tilde{Q}$ as $N \to \infty$, where $\tilde{Q}$ is the unique pathwise solution to*

$$\tilde{Q}(t) = \zeta(t) + \int_0^t \zeta(t-u)dM(u) + \int_0^t \tilde{Q}^-(t-u)dM(u), \tag{59}$$

for $t \geq 0$, where $\tilde{Q}^-(t) = -min(\tilde{Q}(t), 0)$.

Proof. Let $F = \{F(t), t \geq 0\}$ be a distribution function and $r = \{r(t), t \geq 0\}$ be an unknown function satisfying the integral equation of renewal type,

$$r(t) = H(t) + \int_0^t r(t)dF(t-u), \quad \text{for} \quad t \geq 0, \tag{60}$$

for some $H = \{H(t), t \geq 0\}$. If H is a locally bounded function, then (60) has a unique locally bounded solution [24], which is given by

$$r(t) = H(t) + \int_0^t H(t-u)dM(u), \tag{61}$$

where $M = \{M(t), t \geq 0\}$ is the solution to the renewal equation,

$$M(t) = F(t) + \int_0^t M(t-u)dF(u), \quad \text{for} \quad t \geq 0.$$

By the definition of ζ in (58), the limiting process of Theorem 2.4.4 may be written as

$$\tilde{Q}(t) \quad = \quad \zeta(t) + \int_0^t \tilde{Q}^+(s)dF(t-s), \quad \text{for } t \geq 0.$$

Next, since $\tilde{Q} = \tilde{Q}^+ - \tilde{Q}^-$, we have

$$\tilde{Q}^+(t) \quad = \quad \zeta(t) + \tilde{Q}(t)^- + \int_0^t \tilde{Q}^+(s)dF(t-s).$$

Furthermore, it follows that $\zeta + \tilde{Q}^-$ is almost surely a locally bounded function since it is almost surely an element of $D[0, \infty)$. It therefore follows from (60) and (61) that

$$\tilde{Q}^+(t) = \zeta(t) + \tilde{Q}^-(t) + \int_0^t \zeta(t-u)dM(u) - \int_0^t \tilde{Q}^-(t-u)dM(u),$$

or, equivalently,

$$\tilde{Q}(t) = \zeta(t) + \int_0^t \zeta(t-u)dM(u) - \int_0^t \tilde{Q}^-(t-u)dM(u)$$

which completes the proof. □

In the case of the $GI/M/N$ queue, Corollary 2.4.5 may be used to obtain the original diffusion limit result obtained by Halfin and Whitt [15]. This may be seen by first noting that for exponentially distributed service times, the renewal function M in (59) will be the renewal function for a rate 1 Poisson process, which is simply given by $M(t) = t$. Thus, the limiting process of Corollary 2.4.5 may be written

$$\tilde{Q}(t) = \zeta(t) + \int_0^t \zeta(s)ds + \int_0^t \tilde{Q}^-(s)ds. \tag{62}$$

Furthermore, using (58), (48) and (49), extensive covariance calculations show that

$$B(t) \quad = \quad \zeta(t) + \int_0^t \zeta(s)ds, \quad t \geq 0,$$

is a Brownian motion with drift $-\beta$ and infinitesimal variance $1+\sigma^2$, where σ^2 is the variance of the interarrival times. Therefore, the process (62) is a diffusion with infinitesimal drift $m(x) = -\beta$ for $x \geq 0$ and $m(x) = -x - \beta$ for $x < 0$ and infinitesimal variance $1 + \sigma^2$, both in agreement with Theorem 3 of Halfin and Whitt [15].

CHAPTER III

CUSTOMER ABANDONMENT IN HEAVY TRAFFIC

We present a novel heavy traffic regime for the $GI/GI/1+GI$ queue. Our approach is to scale the hazard rate function of the abandonment distribution of the customers arriving to the queue. Our main result is then to provide a limiting diffusion approximation for both the scaled queue length and virtual waiting time processes. The diffusion limits we obtain are reflected at the origin into the positive portion of the real line and have a negative drift which incorporates the entire hazard rate function. The proofs of the lemmas contained in this Chapter may be found in Appendices A and C.

3.1 Model Formulation

Our study of the GI/GI/1 queue having FIFO service and customer abandonments begins with the model introduced in Ward and Glynn [52]. The model primitives are three independent i.i.d. sequences of non-negative random variables $\{u_i, i \geq 1\}, \{v_i, i \geq 1\}$, and $\{a_i, i \geq 1\}$, which are all defined on a common probability space $(\Omega, \mathcal{F}, P)$. We assume that $E[u_1] = E[v_1] = 1$ and $\text{var}(u_1) < \infty, \text{var}(v_1) < \infty$. For a given arrival rate ρ, the ith system arrival joins the queue at time

$$t_i \equiv \sum_{j=1}^{i} \frac{u_j}{\rho},$$

has service time v_i, and will abandon if processing does not begin within a_i time units. (For the interested reader, we note that more sophisticated models of customer impatience can be found in Mandelbaum and Shimkin [38] and Zohar, Mandelbaum, and Shimkin [60].) We let F be the cumulative distribution function of a_1, and

$$h(x) = \frac{\frac{d}{dx}F(x)}{1 - F(x)}, \; x \geq 0$$

be the associated hazard rate function. We assume F is proper; i.e., that $\lim_{x\to\infty} F(x) = 1$. Then, Lemma 2 in Baccelli, Boyer, and Hebuterne [3] guarantees the offered waiting time process given in (63) below possesses a non-degenerate limiting distribution.

The length of time a customer arriving at time t has to wait for service depends upon the processing times of the customers in the queue at time t who eventually receive service (and do not abandon). In particular, for $t > 0$, the *offered waiting time* process

$$V(t) \equiv \sum_{n=1}^{A(t)} v_n \mathbf{1}\{V(t_n^-) < a_n\} - B(t) \geq 0 \tag{63}$$

tracks the waiting time an infinitely patient arriving customer would experience at time $t > 0$. Here, the process

$$A(t) \equiv \max\left\{ i \geq 0 : \sum_{j=1}^{i} u_j \leq \rho t \right\}$$

counts the number of customers that have arrived to the system by time $t \geq 0$ and the process

$$B(t) \equiv \int_0^t \mathbf{1}\{V(s) > 0\} ds$$

is the cumulative server busy time.

It is useful for our analysis to represent the offered waiting time process in terms of a stochastic integral and three martingales as follows. Define the σ-field

$$\mathcal{F}_i \equiv \sigma((u_1, v_1, a_1), ..., (u_i, v_i, a_i), u_{i+1}) \subset \mathcal{F}$$

such that

$$P(V(t_i^-) \geq a_i | \mathcal{F}_{i-1}) = F\left(V(t_i^-)\right), \quad i = 1, 2, \ldots,$$

almost surely, because $V(t_i^-)$ is $\mathcal{F}_{i-1}$ measurable and a_i is independent of $\mathcal{F}_{i-1}$. The martingale

$$\left\{ \left(M_a(i) \equiv \sum_{j=1}^{i} \mathbf{1}\{V(t_j^-) \geq a_j\} - E[\mathbf{1}\{V(t_j^-) \geq a_j\} | \mathcal{F}_{j-1}], \mathcal{F}_i \right), i \geq 0 \right\} \tag{64}$$

is the sum of the random variables representing which customers abandon, centered by their conditional means. Also let

$$S(i) \equiv \sum_{j=1}^{i} (v_j - E[v_1])$$

be the sum of the centered service times and

$$S_a(i) \equiv \sum_{j=1}^{i} (v_j - E[v_1]) \mathbf{1}\{V(t_i^-) \geq a_i\}$$

be the sum of the centered service times of those customers that will eventually abandon. Define the centered process

$$X(t) \equiv E[v_1]A(t) - \rho t + S(A(t)) + t\,(\rho - 1) - S_a(A(t)) - E[v_1]M_a(A(t)), \tag{65}$$

and the "integral error" process

$$\epsilon(t) \equiv \int_0^t \left(\int_0^{V(s^-)} h(u)du \right) ds - E[v_1] \int_0^t F(V(s^-))dA(s). \tag{66}$$

(Note that even though $E[v_1] = 1$, we explicitly write $E[v_1]$ in (65) because its presence will be important in our heavy traffic regime defined in Section 3.2, where service times are scaled to become small.) Algebraic manipulations of (63) show that

$$V(t) = X(t) + \epsilon(t) - \int_0^t \left(\int_0^{V(s^-)} h(u)du \right) ds + I(t), \tag{67}$$

where

$$I(t) \equiv t - B(t) = \int_0^t \mathbf{1}\{V(s) = 0\}ds \tag{68}$$

is the cumulative server idle time.

We first perform our asymptotic analysis under the assumption that the abandonment distribution has support on the positive real line.

Assumption 1. *The abandonment distribution F may be expressed as*

$$F(x) = 1 - \exp\left(- \int_0^x h(u)du \right), \text{ for } x \geq 0$$

where h is a non-negative and continuous function on $[0, \infty)$.

We then extend our analysis to include abandonment distributions having compact support.

Assumption 2. *The abandonment distribution F may be expressed as*

$$F(x) = \left(1 - \exp\left(- \int_0^{x \wedge C} h(u)du \right) \right) + b\mathbf{1}\{x \geq C\},\ x \geq 0,$$

where h is a non-negative and continuous function on $[0, C]$, and $b \equiv \exp\left(- \int_0^C h(u)du \right)$.

Assumption 2 allows distributions such as the deterministic distribution, for which $h(x) = 0,\ x < C$, and $F(x) = \mathbf{1}\{x \geq C\}$, but does not allow distributions such as the uniform distribution on $[0, 1]$, for which $h(x) = (1 - x)^{-1} \to \infty$ as $x \uparrow 1$. For technical reasons, we avoid distributions whose hazard rate tends to infinity on its support.

3.2 Hazard Rate Scaling in Heavy Traffic

We first define our heavy traffic asymptotic regime in Subsection 3.2.1. Subsection 3.2.2 provides intuition for our assumed hazard rate scaling, and Subsection 3.2.3 discusses its implications in terms of customer patience.

3.2.1 The Heavy Traffic Asymptotic Regime

We consider a sequence of systems indexed by $n \geq 1$ in which the arrival rates become large and service times small. Our convention is to superscript any process or quantity associated with the n^{th} system by n. Specifically, the n^{th} system has arrival rate $n\rho^n$. That is, the i^{th} arrival to the n^{th} system occurs at time

$$t_i^n \equiv \sum_{j=1}^{i} \frac{u_j}{n\rho^n},$$

and the cumulative number of customer arrivals in $[0, t]$ in the n^{th} system is given by

$$A^n(t) = \max\{i \geq 0, t_i^n \leq t\}, \quad t \geq 0.$$

The service time of the i^{th} arrival is

$$v_i^n \equiv v_i/n, \tag{69}$$

so that the sum of centered service times becomes

$$S^n(i) = \frac{1}{n} \sum_{j=1}^{i} (v_j - E[v_1]). \tag{70}$$

As n increases, the mean arrival and service rates become arbitrarily close; in particular,

$$\sqrt{n}(\rho^n - 1) \to \theta, \text{ as } n \to \infty, \tag{71}$$

where $\theta \in \Re$.

We scale the hazard rate function by $\sqrt{n}$ so that the hazard rate function associated with customer abandonment times in the n^{th} system is

$$h^n(x) \equiv h(\sqrt{n}x). \tag{72}$$

To intuitively motivate the scaling in (72), first observe that in a conventional queueing system having $a_i = \infty$ for all $i \geq 0$, Kingman's approximation [26] shows the queue size is proportional to $(1-\rho^n)^{-1}$, which is of order $\sqrt{n}$ from assumption (71). Because the arrival rate in the n^{th} system is of order n, a sample path version of Little's law known as the snapshot principle (see Reiman [43]) suggests that

$$V^n(t) \approx \frac{Q^n(t)}{n\rho^n} \propto \sqrt{n}/n = 1/\sqrt{n}, \tag{73}$$

meaning the offered waiting time in the n^{th} system shrinks at rate $n^{-1/2}$ as n grows large. Therefore, as in an observation made much earlier by Lehoczky [34] (and further developed in [10], [31], [32], and [30] for a $GI/GI/1{+}GI$ system operating under the earliest-deadline-first service discipline, and under the assumption that customers do not abandon the system when their deadline expires), in order that the limiting system capture the effects of customer abandonments, customer abandonment times must be shrinking (at rate $\sqrt{n}$) as n grows large. Furthermore, in order that more than only the behavior of the abandonment distribution close to the origin be used to determine whether or not a customer in the n^{th} system abandons when n is large, the hazard rate scaling must inflate its argument by $\sqrt{n}$.

Under Assumption 1, (72) implies the distribution of abandonment times in the n^{th} system is

$$F^n(x) = 1 - \exp\left(-\int_0^x h(\sqrt{n}u)du\right), \text{ for } x \geq 0. \tag{74}$$

Under Assumption 2, (72) implies the upper bound on abandonment times in the n^{th} system is

$$C^n \equiv \frac{C}{\sqrt{n}}, \tag{75}$$

and the distribution of abandonment times in the n^{th} system is

$$F^n(x) = \left(1 - \exp\left(-\frac{1}{\sqrt{n}}\int_0^{(\sqrt{n}x)\wedge C} h(w)dw\right)\right) + b^n \mathbf{1}\{\sqrt{n}x \geq C\}, \tag{76}$$

where

$$b^n \equiv \exp\left(-\frac{1}{\sqrt{n}}\int_0^C h(w)dw\right). \tag{77}$$

We assume customer abandonment times in the n^{th} system are an i.i.d. sequence of random variables $\{a_j^n, j \geq 1\}$ having distribution F^n defined in either (74) or (76). Note that the

sum of centered service times of those customers that will eventually abandon becomes

$$S_a^n(i) = \frac{1}{n}\sum_{j=1}^{i}(v_j - E[v_1])\,\mathbf{1}\{V^n(t_i^{n,-}) \geq a_i^n\}. \tag{78}$$

It is useful for later analysis to define the fluid-scaled quantity

$$\overline{A}^n(t) \equiv \frac{A^n(t)}{n}, \tag{79}$$

and the diffusion-scaled quantities

$$\tilde{V}^n(t) \equiv \sqrt{n}V^n(t) \tag{80}$$

$$\tilde{A}^n(t) \equiv \sqrt{n}\left(\frac{1}{n}A^n(t) - \rho^n t\right) \tag{81}$$

$$\tilde{S}^n(t) \equiv \sqrt{n}S^n(\lfloor nt \rfloor) \tag{82}$$

$$\tilde{S}_a^n(t) \equiv \sqrt{n}S_a^n(\lfloor nt \rfloor) \tag{83}$$

$$\tilde{M}_a^n(t) \equiv \frac{1}{\sqrt{n}}M_a^n(\lfloor nt \rfloor) \tag{84}$$

$$\tilde{I}^n(t) \equiv \sqrt{n}I^n(t). \tag{85}$$

Recall from (73) that the scaling that leads to a non-degenerate limit process should inflate the offered waiting time process by $\sqrt{n}$.

We require the following technicalities. All random variables are defined on a common probability space $(\Omega, \mathcal{F}, P)$. For each positive integer d, let $D([0,\infty), \Re^d)$ be the space of right continuous functions with left limits (RCLL) in $\Re^d$ having time domain $[0,\infty)$. We endow $D([0,\infty), \Re^d)$ with the usual Skorokhod J_1 topology, and let M^d denote the Borel σ-algebra associated with the J_1 topology. All stochastic processes are measurable functions from $(\Omega, \mathcal{F}, P)$ into $(D([0,\infty), \Re^d), M^d)$ for some appropriate dimension d. We will often use the notation $\xi^n = \{\xi^n(t), t \geq 0\}$ to denote the stochastic process associated with a collection of random variables $\{\xi^n(t), t \geq 0\}$. Suppose $\{\xi^n\}_{n=1}^{\infty}$ is a sequence of stochastic processes. The notation $\xi^n \Rightarrow \xi$ means that the probability measures induced by the ξ^n's on $(D([0,\infty), \Re^d), M^d)$ converge weakly to the probability measure on $(D([0,\infty), \Re^d), M^d)$ induced by the stochastic process ξ. The notation $\stackrel{D}{=}$ means equal in distribution.

The functional strong law of large numbers (see, for example, Theorem 5.10 in Chen and Yao [8]) establishes

$$\overline{A}^n \to e, \tag{86}$$

P-almost surely, uniformly on compact sets, as $n \to \infty$, where $e(t) = t$ for all $t \geq 0$ is the identity function. Let $W_{S,1}$ and $W_{S,2}$ be independent, standard Brownian motions. The functional central limit theorem for renewal processes (see, for example Theorem 5.11 in Chen and Yao [8]) establishes

$$\tilde{A}^n \Rightarrow \text{var}(u_1) W_{S,1},$$

as $n \to \infty$, and Donsker's theorem (see, for example, Theorem 14.1 in Billingsley [4]) establishes

$$\tilde{S}^n \Rightarrow \text{var}(v_1) W_{S,2},$$

as $n \to \infty$. The assumed independence of the inter-arrival and service time sequences implies the joint convergence

$$\left(\tilde{A}^n, \tilde{S}^n\right) \Rightarrow \left(\text{var}(u_1) W_{S,1}, \text{var}(v_1) W_{S,2}\right), \tag{87}$$

as $n \to \infty$. We often use the random time change theorem in our proofs, and a convenient statement of this result can be found in Chapter 3, Section 14 of Billingsley [4]. In general, addition is not a continuous map from $D([0,\infty), \Re) \times D([0,\infty), \Re) \to D([0,\infty), \Re)$; however, addition is a continuous map on the subspace of continuous functions. All limit processes in this Chapter are continuous, and so we often use the continuous mapping theorem (see, for example, Theorem 3.4.1 of Whitt [54]) in association with the addition operator and obtained limit processes without further explanation. Finally, the space $D([0,\infty), \Re)$ is separable and complete by Theorem 16.3 in Billingsley [4], and so relative compactness and tightness in $D([0,\infty), \Re)$ are equivalent by Prohorov's theorem (see, for example, Theorem 5.1 in Billingsley [4] for the direct half and Theorem 5.2 in [4] for the converse half). We use the two words interchangeably.

3.2.2 Intuition for the Hazard Rate Scaling

In order to produce an interesting limiting diffusion approximation, we would like the state-dependent rate at which customers are abandoning the system, appropriately scaled, to converge to a non-degenerate limit. To calculate this state-dependent rate, we must first determine the probability that each customer in the queue will abandon the system in

the next small amount of time. We can calculate this abandonment probability for each customer using the hazard rate function associated with the abandonment distribution as follows.

First, as in Whitt [56], we assume that few customers have abandoned, and so the amount of time that the ith customer from the back of the queue has been waiting is close to i/n, because i customers have arrived after this customer, the average inter-arrival time in the nth system is $1/(\rho^n n)$, and ρ^n is close to 1. This then implies that the probability that this customer will abandon the system in the next δ time units is close to $h^n(i/n)\delta$, and so, summing over all the customers in line at time t, the abandonment rate at time t is approximately

$$\sum_{i=1}^{Q^n(t)} h^n\left(\frac{i}{n}\right).$$

From (73), $Q^n \approx n\rho^n V^n$, and so since ρ^n is close to 1 for large n, the above sum is approximately

$$\sum_{i=1}^{nV^n(t)} h^n\left(\frac{i}{n}\right) = \sum_{i=1}^{\sqrt{n}\tilde{V}^n(t)} h\left(\frac{i}{\sqrt{n}}\right),$$

where the equality follows from the definition of h^n in (72) and $\tilde{V}^n$ in (80).

Since the relationship (73) also suggests that the queue-length is growing at rate $\sqrt{n}$, to have any hope of obtaining a finite, state-dependent total system abandonment rate, we must scale by $n^{-1/2}$. Then, assuming that $\tilde{V}^n \Rightarrow V$ as $n \to \infty$, we find that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{\sqrt{n}\tilde{V}^n(t)} h\left(\frac{i}{\sqrt{n}}\right) \Rightarrow \int_0^{V(t)} h(u)du, \tag{88}$$

as $n \to \infty$, also using the definition of the Riemann-Stieltjes integral. Hence we have a limiting regime in which the total system abandonment rate converges to a non-degenerate limit. Moreover, the entire abandonment distribution influences this limiting system abandonment rate.

We further conjecture that for large n, the instantaneous rate of abandonment approximately equals the arrival rate times the probability that an arriving customer abandons. Consider a customer who arrives to the queue in the n^{th} system at time t and finds the offered waiting time to be $V^n(t)$. When abandonment times are unbounded, the probability

that this customer abandons is

$$F^n(V^n(t)) \quad = \quad 1 - \exp\left(- \int_0^{V^n(t)} h(\sqrt{n}u)du \right),$$

and a change of variables shows that the right hand side of the above expression is equivalently rewritten as

$$1 - \exp\left(-\frac{1}{\sqrt{n}} \int_0^{\tilde{V}^n(t)} h(u)du \right).$$

Again assuming $\tilde{V}^n \Rightarrow V$ as $n \to \infty$, a Taylor expansion of e^x about zero shows that the probability an arriving customer abandons decreases at rate $\sqrt{n}$ as n grows large. In particular, applying L'Hopital's rule shows

$$\sqrt{n}F^n(V^n(t)) \Rightarrow \int_0^{V(t)} h(u)du \tag{89}$$

as $n \to \infty$. Comparing the limits (88) and (89) side-by-side shows

$$\sum_{i=1}^{\sqrt{n}\tilde{V}^n(t)} h\left(\frac{i}{\sqrt{n}}\right) \approx nF^n(V^n(t)),$$

suggesting that for large n the rate of abandonment approximately equals the arrival rate times the probability of abandonment.

3.2.3 Implications of the Scaling

In this section, we discuss the implications of our assumed hazard rate scaling (72) on customer impatience. We begin with the following definition. Recall that F^n represents the abandonment time distribution of customers in the n^{th} system.

Definition 3.2.1. We say that customers are becoming more impatient on the diffusive time scale if

$$F^{n+1}((n+1)^{-1/2}x) \geq F^n(n^{-1/2}x),$$

for all $x \in \mathbb{R}$ and $n \geq 1$.

The diffusive time scale is of interest because from (73), customer waiting times in the nth system are of order $n^{-1/2}$.

It is easy to see that customers are always becoming more impatient on the diffusive time scale and we record this as our first proposition of this section. The proof is immediate from the representation for F^n in (74) under Assumption 1 and (76) under Assumption 2, and so is omitted.

Proposition 3.2.2. *Customers are becoming more impatient on the diffusive time scale for any abandonment time distribution satisfying Assumption 1 or 2.*

We next study what is occurring to customer impatience on the original time scale. The following definition characterizes customer impatience in terms of whether the abandonment time distribution function is stochastically increasing or decreasing as n increases.

Definition 3.2.3. We say that customers are becoming more impatient (patient) on the original time scale if $F^{n+1}(x) \geq F^n(x)$ $(F^{n+1}(x) \leq F^n(x))$ for all $x \in \Re$ and $n \geq 1$.

When we do not change the timescale as n increases, the characterization of customer patience levels is more complicated, and requires closer examination of the hazard rate function h. Recall the following definition of a distribution function G with an increasing (decreasing) average hazard rate.

Definition 3.2.4. A distribution function G, with hazard rate h, is said to possess an increasing average hazard rate if for $0 \leq a \leq b$,

$$\frac{1}{a}\int_0^a h(u)du \leq \frac{1}{b}\int_0^b h(u)du.$$

We say that G has a decreasing average hazard rate if the above inequality holds with $b \leq a$.

The following result characterizes impatience on the original time scale.

Proposition 3.2.5. *Under either Assumption 1 or Assumption 2, customers become more impatient (patient) on the original time scale if and only if their abandonment time distribution possesses an increasing (decreasing) average hazard rate.*

Proof of Proposition 3.2.5: Suppose first that F possesses an increasing average hazard rate. Then, for each $x \geq 0$ under Assumption 1 or $0 \leq x < C^n$ under Assumption 2, from the expression for F^n in (74) or (76),

$$1 - F^n(x) = e^{-\frac{1}{\sqrt{n}} \int_0^{\sqrt{n}x} h(u)du} \geq e^{-\frac{1}{\sqrt{n+1}} \int_0^{\sqrt{n+1}x} h(u)du} = 1 - F^{n+1}(x),$$

where the inequality follows since F possesses an increasing average hazard rate.

Suppose, on the other hand, that customers are becoming more impatient on the original time scale. Then, for each $x \geq 0$ under Assumption 1 or $0 \leq x \leq C^n$ under Assumption 2,

$$e^{-\frac{1}{\sqrt{n}} \int_0^{x\sqrt{n}} h(u)du} = 1 - F^n(x) \geq 1 - F^{n+1}(x) = e^{-\frac{1}{\sqrt{n+1}} \int_0^{x\sqrt{n+1}} h(u)du},$$

so that F possess an increasing average hazard rate. □

Because the exponential distribution is the only continuous distribution with a constant hazard rate, we also have the following immediate corollary of Proposition 3.2.5.

Corollary 3.2.6. *The level of customer patience remains constant if and only if the abandonment distribution is exponential.*

3.3 Non-linear Generalized Regulator Mappings

The key to our asymptotic analysis in Section 3.4 (that establishes the weak convergence of the offered waiting time process) is to represent the offered waiting time process in terms of a one- or two-sided non-linear generalized regulator mapping. To see what the appropriate mappings are, first observe that under assumption 1, from (65), (66), (67), and (69), the evolution equation for the offered waiting time process in the n^{th} system is

$$V^n(t) = X^n(t) + \epsilon^n(t) - \int_0^t \left(\int_0^{V^n(s^-)} h^n(u)du \right) ds + I^n(t), \tag{90}$$

where

$$X^n(t) \equiv \frac{1}{n} A^n(t) - \rho^n t + S^n(A^n(t)) + t(\rho^n - 1) - S_a^n(A^n(t)) - \frac{1}{n} M_a\left(A^n(t)\right) \tag{91}$$

and, also using the definition of F^n in (74)

$$\begin{aligned}\epsilon^n(t) &\equiv \int_0^t \left(\int_0^{V^n(s^-)} h^n(u)du\right) ds - \frac{1}{n}\int_0^t F^n\left(V^n(s^-)\right) dA^n(s). \qquad (92)\\ &= \frac{1}{\sqrt{n}}\int_0^t \left(\int_0^{\tilde{V}^n(s^-)} h(w)dw\right) ds \\ &\quad - \int_0^t \left(1-\exp\left(-\frac{1}{\sqrt{n}}\int_0^{\tilde{V}^n(s^-)} h(w)dw\right)\right) d\overline{A}^n(s)\end{aligned}$$

Under Assumption 2, we add and subtract the number of arrivals that find the offered waiting time process exceeding the upper bound on abandonment times, appropriately scaled,

$$U^n(t) \equiv \frac{b^n}{n}\int_0^t \mathbf{1}\{V^n(s^-) \geq C^n\} dA^n(s), \qquad (93)$$

to the right-hand side of (90) to find

$$V^n(t) = X^n(t) + \epsilon_B^n(t) - \int_0^t \left(\int_0^{V^n(s^-)\wedge C^n} h^n(u)du\right) ds + I^n(t) - U^n(t), \qquad (94)$$

where X^n is as defined in (91) and, also using the definition of F^n in (76)

$$\begin{aligned}\epsilon_B^n(t) &\equiv \int_0^t \left(\int_0^{V^n(s^-)\wedge C^n} h^n(u)du\right) ds + \frac{b^n}{n}\int_0^t \mathbf{1}\{V^n(s^-) \geq C^n\} dA^n(s) \qquad (95)\\ &\quad -\frac{1}{n}\int_0^t F^n\left(V^n(s^-)\right) dA^n(s) \\ &= \frac{1}{\sqrt{n}}\int_0^t \left(\int_0^{\tilde{V}^n(s^-)\wedge C} h(w)dw\right) ds \\ &\quad - \int_0^t \left(1-\exp\left(-\frac{1}{\sqrt{n}}\int_0^{\tilde{V}^n(s^-)\wedge C} h(w)dw\right)\right) d\overline{A}^n(s).\end{aligned}$$

Observe that the process I^n in (90) and (94) only increases when V^n is 0 and the process U^n in (93) only increases when V^n is equal to or exceeds C^n. Regarding $X^n + \epsilon^n$ and $X^n + \epsilon_B^n$ as the "free" processes, equations (90) and (94) immediately suggest the non-linear generalizations of the conventional one- and two-sided regulator mappings required to obtain weak convergence results on the offered waiting time process under Assumptions 1 and 2.

The remainder of this section is organized as follows. We define a one-sided non-linear generalized regulator mapping in Subsection 3.3.1, and prove its existence, uniqueness, and

continuity. Next, in Subsection 3.3.2, we do the same for a two-sided non-linear generalized regulator mapping.

3.3.1 The One-Sided Non-Linear Generalized Regulator Mapping

The one-sided non-linear generalized regulator mapping generalizes the conventional one-sided regulator mapping first introduced in Skorokhod [46] having the explicit form

$$\phi(x)(t) \equiv x(t) + \psi(x)(t) \in [0, \infty) \text{ for all } t \geq 0, \tag{96}$$

for $x \in D([0, \infty), \Re)$ and

$$\psi(x)(t) \equiv \sup_{0 \leq s \leq t} \left[-x(s)\right]^+, \tag{97}$$

to a mapping that allows for non-linear state-space dependence.

Definition 3.3.1. (The one-sided nonlinear generalized regulator mapping)
Given h a non-negative, continuous function on $[0, \infty)$ and $x \in D([0, \infty), \Re)$ having $x(0) \geq 0$, the one-sided nonlinear generalized regulator mapping

$$(\phi^h, \psi^h) : D([0, \infty), \Re) \mapsto D([0, \infty), [0, \infty) \times [0, \infty))$$

is defined by

$$(\phi^h, \psi^h)(x) \equiv (z, l)$$

where

(C1) $z(t) = x(t) - \int_0^t \left(\int_0^{z(s)} h(u)du \right) ds + l(t) \in [0, \infty)$ for all $t \geq 0$;

(C2) l is nondecreasing, $l(0) = 0$, and $\int_0^\infty z(t)dl(t) = 0$.

When h is the zero function, ϕ and ψ in (96) and (97) uniquely satisfy Definition 3.3.1. When the function h is constant, the non-linear generalized one-sided regulator mapping becomes the linearly generalized one-sided mapping given in Section 5 of Reed and Ward [42].

For $x \in D([0, \infty), \Re)$ having $x(0) \geq 0$ and (ϕ, ψ) defined in (96) and (97), set

$$z \equiv \phi^h(x) = \phi\left(\mathcal{M}^h(x)\right) \tag{98}$$

$$l \equiv \psi^h(x) = \psi\left(\mathcal{M}^h(x)\right), \tag{99}$$

where the mapping $\mathcal{M}^h : D([0,\infty),\Re) \to D([0,\infty),\Re)$ has $\mathcal{M}^h(x) \equiv w$ for w that solves the integral equation

$$w(t) = x(t) - \int_0^t \left(\int_0^{\phi(w)(s)} h(u)du \right) ds \tag{100}$$

having initial condition $w(0) = x(0)$. Observe that (z, l) defined in (98) and (99) satisfy conditions (C1) and (C2) of Definition 3.3.1 because

1. From (96),(98), (99), and (100),

$$\begin{aligned} 0 &\leq \phi\left(\mathcal{M}^h(x)\right)(t) \\ &= \mathcal{M}^h(x)(t) + \psi\left(\mathcal{M}^h(x)\right)(t) \\ &= x(t) - \int_0^t \left(\int_0^{z(s)} h(u)du \right) ds + l(t); \end{aligned}$$

2. The function $\psi(\mathcal{M}^h(x))$ is non-decreasing from its definition in (97), $\psi\left(\mathcal{M}^h(x)\right)(0) = 0$ since $\mathcal{M}^h(x)(0) = x(0) \geq 0$ by assumption on x, and $\int_0^\infty \phi\left(\mathcal{M}^h(x)\right)(t) d\psi\left(\mathcal{M}^h(x)\right)(t) = 0$ from the definitions of ϕ and ψ in (96) and (97).

Therefore, the key to proving existence, uniqueness, and continuity of the non-linear generalized one-sided regulator mapping in Definition 3.3.1 is the following lemma that establishes the existence, uniqueness, and local Lipschitz continuity of the integral equation in (100). We note that if h is bounded on $[0,\infty)$ then the mapping $\mathcal{M}^h$ is globally Lipschitz; in particular, the constant κ in Lemma 3.3.2, part (ii) below depends only on T and the conclusion in (ii) holds for all $x_1, x_2 \in D([0,\infty),\Re)$.

Lemma 3.3.2. *(Properties of the Integral Equation (100))*
Let h be a non-negative, continuous function on $[0,\infty)$.

(i) For each $x \in D([0,\infty),\Re)$ there exists a unique w satisfying (100).

(ii) Let $T > 0$ and $x \in D([0,\infty),\Re)$. There exists κ, dependent on x, such that if $x_1, x_2 \in D([0,\infty),\Re)$ satisfy

$$\|x_1 - x\|_T < 1 \text{ and } \|x_2 - x\|_T < 1,$$

then

$$\|\mathcal{M}^h(x_1) - \mathcal{M}^h(x_2)\|_T < \kappa \|x_1 - x_2\|_T.$$

(iii) The function $\mathcal{M}^h$ is continuous when the space $D([0,\infty),\Re)$ is endowed with Skorohod J_1 topology.

Our next proposition establishes the existence, uniqueness, and continuity of the non-linear generalized one-sided regulator mapping. It is useful for its proof and also for later analysis to observe that Lemma 13.4.1 of Whitt [54] establishes that for any $x_1, x_2 \in D([0,\infty),\Re)$,

$$\|\psi(x_1) - \psi(x_2)\|_T \leq \|x_1 - x_2\|_T, \tag{101}$$

and, therefore from (96), as in Lemma 13.5.1 of Whitt [54],

$$\|\phi(x_1) - \phi(x_2)\|_T \leq 2\|x_1 - x_2\|_T. \tag{102}$$

Proposition 3.3.3. *(Properties of the Non-linear Generalized One-sided Regulator Mapping)*
Let h be a non-negative, continuous function on $[0,\infty)$.

(i) For each $x \in D([0,\infty),\Re)$ having $x(0) \geq 0$, there exists a unique pair of functions

$$(\phi^h, \psi^h)(x) = (z, l)$$

that satisfies (C1)-(C2) of Definition 3.3.1.

(ii) Suppose $x \in D([0,\infty),\Re)$ and $x(0) \geq 0$. Let $h^n(x) = h(\sqrt{n}x)$ for all $x \geq 0$ be as defined in (72). Then,

$$\sqrt{n}\left(\phi^{h^n}, \psi^{h^n}\right)(x) = \left(\phi^h, \psi^h\right)(\sqrt{n}x).$$

(iii) Let $T > 0$ and $x \in D([0,\infty),\Re)$. There exists κ, dependent on x, such that if $x_1, x_2 \in D([0,\infty),\Re)$ satisfy

$$\|x_1 - x\|_T < 1 \text{ and } \|x_2 - x\|_T < 1,$$

then

$$\|\phi^h(x_1) - \phi^h(x_2)\|_T \vee \|\psi^h(x_1) - \psi^h(x_2)\|_T \leq \kappa\|x_1 - x_2\|_T.$$

(iv) Both the functions ϕ^h and ψ^h are continuous when the space $D([0,\infty),\Re)$ is endowed with the Skorohod J_1 topology.

Proof of (i): Existence follows from the representations (98) and (99) and part (i) of Lemma 3.3.2. To see the representations (98) and (99) are unique, let (z,l) be a solution satisfying (C1)-(C2) of Definition 3.3.1. Because for

$$g(t) \equiv x(t) - \int_0^t \left(\int_0^{z(s)} h(u)du \right) ds,\ t \geq 0, \tag{103}$$

from (C1)

$$z(t) = g(t) + l(t) \geq 0,$$

and l satisfies (C2), we conclude

$$(z,l) \quad = \quad (\phi,\psi)(g). \tag{104}$$

If we now show that $g = \mathcal{M}^h(x)$, we will then have that

$$(z,l) = \left(\phi\left(\mathcal{M}^h(x)\right), \psi\left(\mathcal{M}^h(x)\right) \right)$$

which, by part (i) of Lemma 3.3.2 and the uniqueness of (ϕ,ψ), will uniquely define (z,l). However, by (104) we have $z = \phi(g)$ and so it follows upon substitution into (103) that

$$g(t) = x(t) - \int_0^t \left(\int_0^{\phi(g)(s)} h(u)du \right) ds$$

as desired.

Proof of (ii): Since

$$\left(\phi^{h^n}, \psi^{h^n}\right)(x) = (z,l) \tag{105}$$

satisfies (C1) of Definition 3.3.1,

$$z(t) = x(t) - \int_0^t \left(\int_0^{z(s)} h^n(u)du \right) ds + l(t).$$

Let $z^n \equiv \sqrt{n}z$, $x^n \equiv \sqrt{n}x$, and $l^n \equiv \sqrt{n}l$. Multiply both sides of the above equation by $\sqrt{n}$ to find

$$z^n(t) = x^n(t) - \int_0^t \left(\int_0^{z^n(s)} h(u)du \right) ds + l^n(t).$$

Since also (C2) of Definition 3.3.1 holds for (z, l), $l^n = \sqrt{n}l$ is non-decreasing, $l^n(0) = \sqrt{n}l(0) = 0$, and $\int_0^\infty z^n(t)dl^n(t) = \int_0^\infty nz(t)dl(t) = 0$, we conclude

$$\left(\phi^h, \psi^h\right)(x^n) = (z^n, l^n).$$

Therefore, from the definitions of z^n, l^n, and x^n, and the equality (105),

$$\sqrt{n}\left(\phi^{h^n}, \psi^{h^n}\right)(x) = \sqrt{n}(z, l) = (z^n, l^n) = \left(\phi^h, \psi^h\right)(x^n) = \left(\phi^h, \psi^h\right)(\sqrt{n}x).$$

Proof of (iii): From the representations (98) and (99), the Lipschitz property of ψ and ϕ in (101) and (102), and part (ii) of Lemma 3.3.2,

$$\begin{aligned}
&\|\phi^h(x_1) - \phi^h(x_2)\|_T \vee \|\psi^h(x_1) - \psi^h(x_2)\|_T \\
&= \|\phi\left(\mathcal{M}^h(x_1)\right) - \phi\left(\mathcal{M}^h(x_2)\right)\|_T \vee \|\psi\left(\mathcal{M}^h(x_1)\right) - \psi\left(\mathcal{M}^h(x_2)\right)\|_T \\
&\leq 2\|\mathcal{M}^h(x_1) - \mathcal{M}^h(x_2)\|_T \vee \|\mathcal{M}^h(x_1) - \mathcal{M}^h(x_2)\|_T \\
&\leq 2\kappa\|x_1 - x_2\|_T,
\end{aligned}$$

where κ is as in part (ii) of Lemma 3.3.2.

Proof of (iv): Suppose that $x^n \to x$ as $n \to \infty$ in the Skorohod J_1 topology. Then, from part (iii) of Lemma 3.3.2, we have that

$$\mathcal{M}^h(x^n) \;\to\; \mathcal{M}^h(x) \;\text{ as }\; n \to \infty,$$

in the Skorohod J_1 topology. Thus, since by Theorems 13.4.1 and 13.5.1 in Whitt [54], ϕ and ψ are both continuous in the Skorohod J_1 topology, and compositions of continuous functions are continuous, this then implies that as $n \to \infty$, in the Skorohod J_1 topology,

$$\phi^h(x^n) \;=\; \phi\left(\mathcal{M}^h(x^n)\right) \;\to\; \phi\left(\mathcal{M}^h(x)\right) \;=\; \phi^h(x)$$

and

$$\psi^h(x^n) \;=\; \psi\left(\mathcal{M}^h(x^n)\right) \;\to\; \psi\left(\mathcal{M}^h(x)\right) \;=\; \psi^h(x).$$

□

3.3.2 The Two-Sided Non-Linear Generalized Regulator Mapping

The two-sided non-linear generalized regulator mapping generalizes the conventional two-sided regulator mapping defined in Section 14.8.1 of Whitt [54], and having the explicit form given in Theorem 1.4 in Kruk et al [29].

Definition 3.3.4. Let $C > 0$. Given h a non-negative continuous function on $[0, C]$ and $x \in D([0,\infty), \Re)$ having $0 \leq x(0) \leq C$, the two-sided non-linear generalized regulator mapping

$$\left(\phi_C^h, \psi_{1,C}^h, \psi_{2,C}^h\right) : D\left([0,\infty), \Re\right) \to D\left([0,\infty), [0,C] \times [0,\infty) \times [0,\infty)\right)$$

is defined by

$$\left(\phi_C^h, \psi_{1,C}^h, \psi_{2,C}^h\right) \equiv (z, l, u)$$

where

(C1) $z(t) = x(t) - \int_0^t \left(\int_0^{z(s)} h(u)du\right) ds + l(t) - u(t) \in [0, C]$ for all $t \geq 0$;

(C2) l and u are non-decreasing, $l(0) = u(0) = 0$, and $\int_0^\infty z(t)dl(t) = \int_0^\infty [C - z(t)]^+ du(t) = 0$.

Similar to Section 3.3.1, when h is the zero function, Definition 3.3.4 defines the conventional two-sided regulator mapping, and we denote the unique mapping by $(\phi_C, \psi_{1,C}, \psi_{2,C})$. When the function h is constant, the non-linear generalized two-sided regulator mapping becomes the linearly generalized two-sided mapping given in Definition 2 of Ward and Kumar [53].

For $x \in D([0,\infty), \Re)$ having $0 \leq x(0) \leq C$, set

$$z \equiv \phi_C^h(x) = \phi_C\left(\mathcal{M}_C^h(x)\right) \tag{106}$$

$$l \equiv \psi_{1,C}^h(x) = \psi_{1,C}\left(\mathcal{M}_C^h(x)\right) \tag{107}$$

$$u \equiv \psi_{2,C}^h(x) = \psi_{2,C}\left(\mathcal{M}_C^h(x)\right), \tag{108}$$

where the mapping $\mathcal{M}_C^h : D([0,\infty), \Re) \to D([0,\infty), \Re)$ has $\mathcal{M}_C^h(x) \equiv w$ for w that solves the integral equation

$$w(t) = x(t) - \int_0^t \left(\int_0^{\phi_C(w)(s)} h(u)du\right) ds \tag{109}$$

having initial condition $w(0) = x(0)$. By paralleling the arguments in the beginning of Section 3.3.1, it is straightforward to show that (z, l, u) defined in (106)-(108) satisfy conditions (C1) and (C2) of Definition 3.3.4. Therefore, the key to understanding the properties of the nonlinear generalized two-sided regulator mapping in Definition 3.3.4 is to understand the properties of the integral equation (109).

Lemma 3.3.5. *(Properties of the Integral Equation (109))*

Let h be a non-negative, continuous function on $[0, C]$.

(i) For each $x \in D([0,\infty), \Re)$, there exists a unique w satisfying (109).

(ii) Let $T > 0$. There exists a finite constant κ that depends only on T such that for any $x_1, x_2 \in D([0,\infty), \Re)$ having $0 \le x_1(0), x_2(0) \le C$,

$$\|\mathcal{M}_C^h(x_1) - \mathcal{M}_C^h(x_2)\|_T \le \kappa \|x_1 - x_2\|_T.$$

(iii) The function $\mathcal{M}_C^h$ is continuous when the space $D([0,\infty), \Re)$ is endowed with the Skorohod J_1 topology.

The main proposition of this subsection establishes several useful properties of the nonlinear generalized two-sided regulator mapping.

Proposition 3.3.6. *(Properties of the Non-linear Generalized Two-Sided Regulator Mapping)*

Let h be a non-negative, continuous function on $[0, C]$.

(i) For each $x \in D([0,\infty), \Re)$ having $0 \le x(0) \le C$, there exists a unique pair of functions

$$\left(\phi_C^h, \psi_{1,C}^h, \psi_{2,C}^h\right)(x) = (z, l, u)$$

that satisfies (C1)-(C2) of Definition 3.3.4.

(ii) Suppose $x \in D([0,\infty), \Re)$ and $0 \le x(0) \le C$. Let $h^n(x) = h(\sqrt{n}x)$ for all $x \ge 0$ be as defined in (72). Then,

$$\sqrt{n}\left(\phi_C^{h^n}, \psi_{1,C}^{h^n}, \psi_{2,C}^{h^n}\right)(x) = \left(\phi_{\sqrt{n}C}^h, \psi_{1,\sqrt{n}C}^h, \psi_{2,\sqrt{n}C}^h\right)(\sqrt{n}x).$$

(iii) Let $T > 0$. There exists a finite constant κ that depends only on T such that for any $x_1, x_2 \in D([0,\infty), \Re)$ having $0 \leq x_1(0), x_2(0) \leq C$

$$\|\phi_C^h(x_1) - \phi_C^h(x_2)\|_T \leq \kappa \|x_1 - x_2\|_T.$$

Furthermore[1], if $\|x^n - x\|_T \to 0$ as $n \to \infty$, then

$$\|\psi_{j,C}^h(x^n) - \psi_{j,C}^h(x)\|_T \to 0,\ j \in \{1,2\}$$

as $n \to \infty$.

(iv) The functions $\phi_C^h, \psi_{1,C}^h$, and $\psi_{2,C}^h$ are continuous when the space $D([0,\infty), \Re)$ is endowed with the Skorohod J_1 topology.

Proof of (i): Existence follows from the representations (106)-(108) and part (i) of Lemma 3.3.5. The proof of uniqueness is very similar to part (i) of Proposition 3.3.3, and so is omitted.

Proof of (ii): Since

$$\left(\phi_C^{h^n}, \psi_{1,C}^{h^n}, \psi_{2,C}^{h^n}\right)(x) = (z, l, u) \tag{110}$$

satisfies (C1) of Definition 3.3.4,

$$z(t) = x(t) - \int_0^t \left(\int_0^{z(s)} h^n(u)du\right) ds + l(t) - u(t) \in [0, C].$$

Let $z^n \equiv \sqrt{n}z$, $x^n \equiv \sqrt{n}x$, $l^n \equiv \sqrt{n}l$, $u^n \equiv \sqrt{n}u$, and $C^n \equiv \sqrt{n}C$. Multiply both sides of the above equation by $\sqrt{n}$ to find

$$z^n(t) = x^n(t) - \int_0^t \left(\int_0^{z^n(s)} h(w)dw\right) ds + l^n(t) - u^n(t) \in [0, C^n].$$

Since also (C2) of Definition 3.3.4 holds for (z, l, u), l^n and u^n are non-decreasing, $l^n(0) = u^n(0) = 0$, and

$$\int_0^\infty z^n(t)dl^n(t) = \int_0^\infty nz(t)dl(t) = 0$$
$$\int_0^\infty [C^n - z^n(t)]^+ du^n(t) = \int_0^\infty n[C - z(t)]^+ du(t) = 0,$$

[1] *We remark that the Lipschitz property does not hold for $\psi_{1,C}^h$ and $\psi_{2,C}^h$. See Example 14.8.1 of Whitt [54] for a counterexample to the Lipschitz property for the conventional two-sided regulator mapping.*

we conclude

$$\left(\phi^h_{C^n}, \psi^h_{1,C^n}, \psi^h_{2,C^n}\right)(x^n) = (z^n, l^n, u^n)\,.$$

Therefore, from the definitions of z^n, l^n, and x^n, and the equality (110),

$$\sqrt{n}\left(\phi^{h^n}_C, \psi^{h^n}_{1,C}, \psi^{h^n}_{2,C}\right)(x) = \sqrt{n}(z,l,u) = (z^n, l^n, u^n) = \left(\phi^h_{\sqrt{n}C}, \psi^h_{1,\sqrt{n}C}, \psi^h_{2,\sqrt{n}C}\right)(\sqrt{n}x).$$

Proof of (iii): From the representations (106), the Lipschitz property of $\phi_{[0,C]}$ established in Theorem 14.8.1 of [54] with Lipschitz constant 2, and part (i) of Lemma 3.3.5,

$$\begin{aligned}
\|\phi^h_C(x_1) - \phi^h_C(x_2)\|_T &= \|\phi_{[0,C]}\left(\mathcal{M}^h_C(x_1)\right) - \phi_{[0,C]}\left(\mathcal{M}^h_C(x_2)\right)\|_T \\
&\leq 2\|\mathcal{M}^h_C(x_1) - \mathcal{M}^h_C(x_2)\|_T \\
&\leq 2\kappa\|x_1 - x_2\|_T.
\end{aligned}$$

Next, assume $\|x^n - x\|_T \to 0$ as $n \to \infty$. Then, part (ii) of Lemma 3.3.5 guarantees

$$\|\mathcal{M}^h_C(x^n) - \mathcal{M}^h_C(x)\|_T \leq \kappa\|x^n - x\|_T \to 0,$$

as $n \to \infty$. The representations of $\psi^h_{1,C}$ and $\psi^h_{2,C}$ in (107) and (108) and the continuity of the mappings $\psi_{1,C}$ and $\psi_{2,C}$ established in Theorem 14.8.1 in Whitt [54] then show that since a composition of continuous functions is continuous

$$\|\psi^h_{j,C}(x^n) - \psi^h_{j,C}(x)\|_T = \|\psi_{j,C}\left(\mathcal{M}^h_C(x^n)\right) - \psi_{j,C}\left(\mathcal{M}^h_C(x)\right)\|_T \to 0,$$

as $n \to \infty$.

Proof of (iv): Since by Theorem 14.8.2 of Whitt [54], $\phi_C, \psi_{1,C}$, and $\psi_{2,C}$ are all continuous in the Skorohod J_1 topology, by part (iv) of Lemma 3.3.5, the proof now proceeds in the same manner as the proof of part (iii) of Proposition 3.3.3. □

3.4 Weak Convergence of the Offered Waiting Time Process

We establish the weak convergence of the scaled offered waiting time process $\tilde{V}^n$ in (80) when the abandonment distribution has unbounded support (Assumption 1) in Section 3.4.1, and

when the abandonment distribution has bounded support (Assumption 2) in Section 3.4.2. Specifically, we prove the following theorem.

Theorem 3.4.1. *Let W be a Brownian motion with drift θ given in (71), variance $\sigma^2 = var(u_1) + var(v_1)$, and initial position $W(0) = 0$.*

(i) Under assumption 1, $\left(\tilde{V}^n, \tilde{I}^n\right) \Rightarrow \left(\phi^h, \psi^h\right)(W)$, as $n \to \infty$.

(ii) Under assumption 2, $\left(\tilde{V}^n, \tilde{I}^n, \tilde{U}^n\right) \Rightarrow \left(\phi^h_C, \psi^h_{1,C}, \psi^h_{2,C}\right)(W)$, as $n \to \infty$.

The limiting virtual waiting time process of Theorem 3.4.1 may loosely be described as a diffusion process with infinitesimal drift given by

$$m(x) = \theta x - \int_0^x h(u)du, \quad \text{for } x \geq 0,$$

and infinitesimal variance σ^2. There is a lower reflecting barrier at the origin for the case of part (i). Part (ii) also requires an upper reflecting barrier at the point C which represents an upper limit on the abandonment times.

Define

$$\tilde{X}^n(t) \quad = \quad \sqrt{n}X^n(t) \tag{111}$$

and

$$\tilde{\epsilon}^n(t) \quad = \quad \sqrt{n}\epsilon^n(t). \tag{112}$$

3.4.1 Proof of Theorem 3.4.1 part (i):

The key to our weak convergence proof is to represent the offered waiting time process in terms of the one-sided non-linear generalized regulator mapping

$$(V^n, I^n) = \left(\phi^{h^n}, \psi^{h^n}\right)(X^n + \epsilon^n), \tag{113}$$

from which the representation of the scaled offered waiting time process in terms of (ϕ^h, ψ^h) follows. To see that (113) is valid, first observe that the evolution equation (90) combined with the original definition of V^n in (63) that guarantees $V^n(t) \geq 0$ for all $t \geq 0$ implies

condition (C1) of Definition 3.3.1 holds. Next, from (68), for every n, I^n is non-decreasing, has $I^n(0) = 0$, and

$$\int_0^\infty V^n(t)dI^n(t) = \int_0^\infty V^n(t)\mathbf{1}\{V^n(t) = 0\}dt = 0,$$

and so (C2) is also satisfied.

The definitions of $\tilde{V}^n$ and $\tilde{I}^n$ in (80) and (85), the representation (113), the scaling property of the non-linear generalized one-sided regulator mapping in part (ii) of Proposition 3.3.3, and the definitions of $\tilde{X}^n$ in (111) and $\tilde{\epsilon}^n$ in (112) imply

$$\begin{aligned}\left(\tilde{V}^n, \tilde{I}^n\right) &= \sqrt{n}\,(V^n, I^n) \\ &= \sqrt{n}\left(\phi^{h^n}, \psi^{h^n}\right)(X^n + \epsilon^n) \\ &= \left(\phi^h, \psi^h\right)\left(\sqrt{n}\,(X^n + \epsilon^n)\right) \\ &= \left(\phi^h, \psi^h\right)\left(\tilde{X}^n + \tilde{\epsilon}^n\right).\end{aligned} \tag{114}$$

Suppose we can show

$$\tilde{X}^n \Rightarrow W \text{ and } \tilde{\epsilon}^n \Rightarrow 0, \tag{115}$$

as $n \to \infty$. Then, the continuous mapping theorem establishes

$$\tilde{X}^n + \tilde{\epsilon}^n \Rightarrow W,$$

as $n \to \infty$. The result in part (i) then follows from the representation of $\left(\tilde{V}^n, \tilde{I}^n\right)$ in (114), the continuous mapping theorem, and part (iv) of Proposition 3.3.3.

To show (115), we require the following three lemmas. The first establishes that the process

$$R^n(i) \equiv \sum_{j=1}^{i} \mathbf{1}\{V^n\left(t_j^{n,-}\right) \geq a_j^n\}, \tag{116}$$

defined so that $R^n\left(A^n(t)\right)$ is the cumulative number of customers in $[0, t]$ who have arrived by time t and will either already abandoned or will abandon after time t, is small on fluid-scale. Define

$$\overline{R}^n(t) \equiv \frac{1}{n} R^n\left(\lfloor nt \rfloor\right). \tag{117}$$

Lemma 3.4.2. *Under assumption 1, as $n \to \infty$, $\overline{R}^n \Rightarrow 0$.*

The second establishes the weak convergence of the diffusion-scaled martingale $\tilde{M}_a^n$ in (84) to the zero process.

Lemma 3.4.3. *Under assumption 1, as $n \to \infty$, $\tilde{M}_a^n \Rightarrow 0$.*

The third establishes tightness of the offered waiting time process.

Lemma 3.4.4. *The sequence $\{\tilde{V}^n\}$ is tight in $D([0,\infty), \Re)$.*

3.4.1.1 Weak convergence of $\tilde{X}^n$ in (115):

From the definition of $\tilde{X}^n$ in (111), the evolution equation for X^n in (91), and the diffusion-scaled processes definitions in (81), (82), (83), and (84)

$$\tilde{X}^n(t) = \tilde{A}^n(t) + \tilde{S}^n\left(\overline{A}^n(t)\right) + \sqrt{n}t\left(\rho^n - 1\right) - \tilde{S}_a^n\left(\overline{A}^n(t)\right) - \tilde{M}_a^n\left(\overline{A}^n(t)\right). \tag{118}$$

Because the service time sequence is i.i.d., recalling the definitions of S_a^n and $\tilde{S}_a^n$ in (78) and (83) and R^n in (116), for any $t \geq 0$,

$$\tilde{S}_a^n\left(\overline{A}^n(t)\right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{A^n(t)} \left(v_j - E\left[v_1\right]\right) \mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n\right\} \stackrel{D}{=} \frac{1}{\sqrt{n}} \sum_{j=1}^{R^n(A^n(t))} \left(w_j - E\left[w_1\right]\right),$$

where $\{w_i, i \geq 1\}$ is an i.i.d. sequence of random variables with distribution equal to that of the service time distribution and which is also independent of the model primitives introduced in Section 3.1. It is straightforward to show that the finite-dimensional distributions are also equivalent, and so the definitions of S^n, $\tilde{S}^n$, and $\overline{R}^n$ in (70), (82), and (117) imply

$$\tilde{S}^n \circ \overline{R}^n \circ \overline{A}^n \stackrel{D}{=} \tilde{S}_a^n \circ \overline{A}^n. \tag{119}$$

The almost sure convergence of $\overline{A}^n$ in (86), the weak convergence of $\overline{R}^n$ in Lemma 3.4.2, the weak convergence of $\tilde{S}^n$ in (87), and the random time change theorem show $\tilde{S}^n \circ \overline{R}^n \circ \overline{A}^n \Rightarrow 0$, as $n \to \infty$, and so the distributional equality in (119) implies

$$\tilde{S}_a^n \Rightarrow 0, \tag{120}$$

as $n \to \infty$. Finally, the representation of $\tilde{X}^n$ in (118), the almost sure convergence of $\overline{A}^n$ in (86), the weak convergences in (87) and (120), the heavy traffic assumption (71), Lemma 3.4.3, and the random time change theorem imply

$$\tilde{X}^n \Rightarrow W,$$

as $n \to \infty$.

3.4.1.2 Weak convergence of $\tilde{\epsilon}^n$ in (115):

Let $\{n_k\}$ be a subsequence along which

$$\tilde{V}^{n_k} \Rightarrow V,$$

as $n_k \to \infty$. Such a subsequence exists because $\{\tilde{V}^n\}$ is tight in $D([0,\infty),\Re)$ by Lemma 3.4.4. From (86), because $\overline{A}^n$ converges to a deterministic limit process, the joint convergence

$$\left(\tilde{V}^{n_k}, \overline{A}^{n_k}\right) \Rightarrow (V, e)\,,$$

as $n_k \to \infty$, is valid. The Skorokhod representation theorem (see, for example, Theorem 3.2.2 in Whitt [54]) guarantees there exists additional random elements on $(D([0,\infty),\Re), J_1) \times D([0,\infty),\Re), J_1)$, $\left\{\left(\breve{V}^{n_k}, \breve{A}^{n_k}\right)\right\}$ and $\breve{V}$, defined on a possibly additional probability space $(\breve{\Omega}, \breve{\mathcal{F}}, \breve{P})$ such that

$$\left(\breve{V}^{n_k}, \breve{A}^{n_k}\right) \stackrel{D}{=} \left(\tilde{V}^{n_k}, \overline{A}^{n_k}\right), \quad \breve{V} \stackrel{D}{=} V, \tag{121}$$

and

$$\breve{P}\left(\lim_{n_k \to \infty} \left(\breve{V}^{n_k}, \breve{A}^{n_k}\right) = \left(\breve{V}, e\right)\right) = 1. \tag{122}$$

Define

$$\begin{aligned} \breve{\epsilon}^n(t) \equiv & \int_0^t \left(\int_0^{\breve{V}^n(s^-)} h(w)dw - \int_0^{\breve{V}(s^-)} h(w)dw \right) ds \\ & + \int_0^t \left(\int_0^{\breve{V}(s^-)} h(w)dw \right) ds \\ & - \int_0^t \sqrt{n} \left(1 - \exp\left(-\frac{1}{\sqrt{n}} \int_0^{\breve{V}^n(s^-)} h(w)dw \right) \right) d\breve{A}^n(s). \end{aligned} \tag{123}$$

Observe from the definition of $\tilde{\epsilon}^n$ in (92), ϵ^n in (66), and h^n in (72) that

$$\begin{aligned} \tilde{\epsilon}^n(t) &= \sqrt{n}\epsilon^n(t) \\ &= \int_0^t \left(\int_0^{\tilde{V}^n(s^-)} h(w)dw \right) ds \\ &\quad - \int_0^t \sqrt{n} \left(1 - \exp\left(-\frac{1}{\sqrt{n}} \int_0^{\tilde{V}^n(s^-)} h(w)dw \right) \right) d\overline{A}^n(s). \end{aligned} \tag{124}$$

From the distributional equivalence (121), the definition of $\breve{\epsilon}^n$ in (123), and the representation of $\tilde{\epsilon}^n$ in (124),

$$\tilde{\epsilon}^n \stackrel{D}{=} \breve{\epsilon}^n. \tag{125}$$

We now show that

$$\breve{\epsilon}^{n_k} \Rightarrow 0, \tag{126}$$

as $n_k \to \infty$. From the continuity of the integrand operator and the convergence in (122),

$$\int_0^{\breve{V}^{n_k}(\cdot^-)} h(w)dw \to \int_0^{\breve{V}(\cdot^-)} h(w)dw,$$

almost surely, uniformly on compact sets of $[0,\infty)$, as $n_k \to \infty$, and so

$$\int_0^{\cdot} \left(\int_0^{\breve{V}^{n_k}(s^-)} h(w)dw - \int_0^{\breve{V}(s^-)} h(w)dw \right) ds \to 0,$$

almost surely, uniformly on compact sets of $[0,\infty)$, as $n_k \to \infty$. Since

$$\sqrt{n}\left(1 - \exp\left(\frac{-x}{\sqrt{n}}\right)\right) \to x,$$

as $n \to \infty$, uniformly on compact sets, we find

$$\sqrt{n_k}\left(1 - \exp\left(\frac{-\int_0^{\breve{V}^{n_k}(\cdot^-)} h(w)dw}{\sqrt{n_k}}\right)\right) \to \int_0^{\breve{V}(\cdot^-)} h(w)dw,$$

almost surely, uniformly on compact sets of $[0,\infty)$, as $n_k \to \infty$. Lemma 8.3 in Dai and Dai [9] and (122) then shows

$$\int_0^{\cdot} \left(\int_0^{\breve{V}_k(s^-)} h(w)dw \right) ds - \int_0^{\cdot} \sqrt{n_k}\left(1 - \exp\left(\frac{\int_0^{\breve{V}^{n_k}(s^-)} h(w)dw}{\sqrt{n}}\right)\right) d\breve{A}^{n_k}(s) \to 0,$$

almost surely, uniformly on compact sets of $[0,\infty)$ as $n_k \to \infty$. We conclude from (123) that the weak convergence in (126) holds.

The distributional equivalence in (125) then implies $\tilde{\epsilon}^{n_k} \Rightarrow 0$ as $n_k \to \infty$. Since the choice of subsequence $\{n_k\}$ was arbitrary, we conclude

$$\tilde{\epsilon}^n \Rightarrow 0,$$

as $n \to \infty$. □

3.4.2 Weak Convergence under Assumption 2

We desire to represent the offered waiting time process in (94) in terms of the non-linear generalized two-sided regulator mapping. However, because V^n may sometimes exceed C^n (which is easily seen from the evolution equation (63)), we cannot directly represent V^n using the two-sided non-linear generalized regulator mapping. Instead, we introduce the process

$$\mathcal{V}^n(t) \equiv V^n(t) \wedge C^n \text{ for all } t \geq 0, \tag{127}$$

and observe that

$$V^n(t) = \mathcal{V}^n(t) + \delta^n(t), \tag{128}$$

where

$$\delta^n(t) \equiv [V^n(t) - C^n]^+ . \tag{129}$$

The following lemma shows that V^n exceeds C^n less and less often and by smaller and smaller amounts in our heavy traffic asymptotic regime. Therefore, representing the process $\mathcal{V}^n$ in terms of the two-sided non-linear generalized regulator mapping allows us to use the same continuous mapping strategy as in our proof of part (i) in Subsection 3.4.1 to obtain weak convergence results for the process V^n. Let

$$\tilde{\delta}^n(t) = \sqrt{n}\delta^n(t). \tag{130}$$

Lemma 3.4.5. *Under assumption 2, as $n \to \infty$, $\tilde{\delta}^n \Rightarrow 0$.*

The processes $\mathcal{V}^n$, I^n, and U^n can be represented as follows

$$(\mathcal{V}^n, I^n, U^n) = \left(\phi^{h^n}_{C^n}, \psi^{h^n}_{1,C^n}, \psi^{h^n}_{2,C^n}\right)(X^n + \epsilon^n_B - \delta^n) . \tag{131}$$

To see (131) is valid, first observe from (94), (127), (128), and because the process V^n is non-negative that $0 \leq \mathcal{V}^n(t) \leq C^n$ for all $t \geq 0$ and

$$\mathcal{V}^n(t) = (X^n(t) + \epsilon^n_B(t) - \delta^n(t)) - \int_0^t \left(\int_0^{\mathcal{V}^n(s^-)} h^n(u)du \right) ds + I^n(t) - U^n(t),$$

meaning condition (C1) of Definition 3.3.4 holds. Next, from the definitions of I^n in (68) and U^n in (93), for every n, I^n and U^n are non-decreasing, have $I^n(0) = U^n(0) = 0$, and

$$\int_0^\infty \mathcal{V}^n(t) dI^n(t) = \int_0^\infty (V^n(t) \wedge C^n) \, \mathbf{1}\{V^n(t) = 0\} dt = 0$$
$$\int_0^\infty [C^n - \mathcal{V}^n(t)]^+ \, dU^n(t) = \frac{b^n}{n} \int_0^\infty [C^n - (V^n(t) \wedge C^n)]^+ \, \mathbf{1}\{V^n(t^-) \geq C^n\} dA^n(t) = 0,$$

and so condition (C2) of Definition 3.3.4 also holds.

Although we cannot directly parallel the representation (114) in the proof of part (i) in Subsection 3.4.1, we can use the non-linear generalized two-sided regulator mapping and the scaled processes

$$\tilde{\mathcal{V}}^n(t) = \sqrt{n}\mathcal{V}^n(t) \tag{132}$$
$$\tilde{\epsilon}_B^n(t) = \sqrt{n}\epsilon_B^n(t) \tag{133}$$
$$\tilde{U}^n(t) = \sqrt{n}U^n(t) \tag{134}$$

to establish a representation for $\left(\tilde{V}^n, \tilde{I}^n, \tilde{U}^n\right)$ that is similar in spirit. First observe from the scalings for $\tilde{\mathcal{V}}^n$, $\tilde{I}^n$, and $\tilde{U}^n$ in (132), (85), and (134), the representation of $(\mathcal{V}^n, I^n, U^n)$ in terms of the non-linear generalized two-sided regulator mapping in (131), the definition of C^n in (75), and the scaling property of the non-linear generalized two-sided regulator mapping in part (ii) of Proposition 3.3.6 that

$$\begin{aligned} \left(\tilde{\mathcal{V}}^n, \tilde{I}^n, \tilde{U}^n\right) &= \sqrt{n}\left(\phi_{C^n}^{h^n}, \psi_{1,C^n}^{h^n}, \psi_{2,C^n}^{h^n}\right)(X^n + \epsilon_B^n - \delta^n) \\ &= \left(\phi_C^h, \psi_{1,C}^h, \psi_{2,C}^h\right)\left(\tilde{X}^n + \tilde{\epsilon}_B^n - \tilde{\delta}^n\right), \end{aligned} \tag{135}$$

also recalling the scalings for $\tilde{X}^n$, $\tilde{\epsilon}_B^n$, and $\tilde{\delta}^n$ in (111), (133), and (130). The representation for V^n in terms of $\mathcal{V}^n$ and δ^n in (128) then implies

$$\left(\tilde{V}^n, \tilde{I}^n, \tilde{U}^n\right) = \left(\tilde{\mathcal{V}}^n, \tilde{I}^n, \tilde{U}^n\right) + \left(\tilde{\delta}^n, 0, 0\right). \tag{136}$$

We parallel the proof of part (i) in Subsection 3.4.1 to show

$$\left(\tilde{\mathcal{V}}^n, \tilde{I}^n, \tilde{U}^n\right) \Rightarrow \left(\phi_C^h, \psi_{1,C}^h, \psi_{2,C}^h\right)(W), \tag{137}$$

as $n \to \infty$. Lemma 3.4.5, the continuous mapping theorem, and the equality (136) then establish the result stated in part (ii) of Theorem 3.4.1.

From the representation (135), Lemma 3.4.5, the continuous mapping theorem, and part (iv) of Proposition 3.3.6, establishing

$$\tilde{X}^n \Rightarrow W \text{ and } \tilde{\epsilon}_B^n \Rightarrow 0, \tag{138}$$

as $n \to \infty$, is sufficient to show (137). We require the following three Lemmas, which are the equivalents of Lemmas 3.4.2-3.4.4 when abandonment times are bounded.

Lemma 3.4.6. *Under assumption 2, as* $n \to \infty$, $\overline{R}^n \Rightarrow 0$.

Lemma 3.4.7. *Under assumption 2, as* $n \to \infty$, $\tilde{M}_a^n \Rightarrow 0$.

Lemma 3.4.8. *The sequence* $\{\tilde{\mathcal{V}}^n\}$ *is tight in* $D([0,\infty), \Re)$.

By Lemmas 3.4.6 and 3.4.7, the arguments showing $\tilde{X}^n \Rightarrow W$ as $n \to \infty$ in Subsection 3.4.1.1 remain valid. The definitions of ϵ_B^n and $\tilde{\epsilon}_B^n$ in (95) and (133), the definitions of $\mathcal{V}^n$ and $\tilde{\mathcal{V}}^n$ in (127) and (132), and the representation of F^n in (76) show

$$\tilde{\epsilon}_B^n(t) = \int_0^t \left(\int_0^{\tilde{\mathcal{V}}^n(s^-)} h(w)dw \right) ds - \int_0^t \sqrt{n} \left(1 - \exp\left(-\frac{1}{\sqrt{n}} \int_0^{\tilde{\mathcal{V}}^n(s^-)} h(w)dw \right) \right) d\overline{A}^n(s). \tag{139}$$

The representation of $\tilde{\epsilon}_B^n$ in (139) above has exactly the same form as that for $\tilde{\epsilon}^n$ in (124) in the proof of part (i) in Subsection 3.4.1, with $\tilde{\mathcal{V}}^n(s^-)$ replacing $\tilde{V}^n(s^-)$. Therefore, because Lemma 3.4.8 establishes the sequence $\{\tilde{\mathcal{V}}^n\}$ is tight in $D([0,\infty), \Re)$, the arguments in Subsection 3.4.1.2 showing $\tilde{\epsilon}^n \Rightarrow 0$ as $n \to \infty$ also show $\tilde{\epsilon}_B^n \Rightarrow 0$, as $n \to \infty$. □

3.5 Stationary Performance Measure Approximation

We first show in Subsection 3.5.1 that the asymptotic behavior of the diffusion-scaled queue-length and offered waiting time processes are identical. Next, in Subsection 3.5.2, we derive the stationary distributions of the limiting diffusion processes in Theorem 3.4.1, which can be used to approximate steady-state performance measures for a GI/GI/1 queue with abandonments. Finally, we perform a simulation study in Subsection 3.5.3 to evaluate the accuracy of our proposed steady-state performance measure approximations.

3.5.1 An Asymptotic Relationship Between the Queue-length and Offered Waiting Time Processes

We establish an asymptotic relationship between the queue-length and offered waiting time processes identical to that in Theorem 4 in Section 3 in Reiman [44] for a conventional GI/GI/1 queue. To handle the complications imposed by the presence of customers that may abandon the system before receiving service, we require the following Lemma. For $t \geq 0$, let $a^n(t)$ be the arrival time of the customer in service at time t in the nth system. If the server is idle, let $a^n(t) = t$.

Lemma 3.5.1. *Under either assumption 1 or 2,*

$$n^{-1/2} \sum_{i=A^n \circ a^n(\cdot)}^{A^n(\cdot)} \mathbf{1}\{V^n\left(t_i^{n,-}\right) \geq a_i^n\} \Rightarrow 0,$$

as $n \to \infty$.

Let $Q^n(t)$ be the queue-length at time $t \geq 0$ in the n^{th} system, and $\tilde{Q}^n(t) = n^{-1/2}Q^n(t)$ be the diffusion-scaled queue-length.

Theorem 3.5.2. *Under either assumption 1 or 2,*

$$\tilde{Q}^n - \tilde{V}^n \Rightarrow 0,$$

as $n \to \infty$.

For the proofs of both Lemma 3.5.1 and Theorem 3.5.2, it is useful to notice that the convergence in (16) of Theorem 4 in Section 3 in [44] continues to hold in our setting. In particular, because the server works at rate 1 and the system is FIFO,

$$V^n(a^n(t)^-) \leq t - a^n(t) \leq V^n(a^n(t)^-) + v^n_{A^n(a^n(t))},$$

and so, recalling the scaling of the service times in (69) and the definition of $\tilde{V}^n$ in (80),

$$\tilde{V}^n\left(a^n(t)^-\right) \leq \sqrt{n}\left(t - a^n(t)\right) \leq \tilde{V}^n\left(a^n(t)^-\right) + \frac{v_{A^n(a^n(t))}}{\sqrt{n}}.$$

Because $\sup_{k=1,\ldots,nt} n^{-1/2} v_k \Rightarrow 0$ as $n \to \infty$ from Lemma 3 in Iglehart and Whitt [20], for each $T \geq 0$,

$$\sup_{0 \leq t \leq T} \left| \sqrt{n}\left(t - a^n(t)\right) - \tilde{V}^n\left(a^n(t)^-\right) \right| \Rightarrow 0, \tag{140}$$

as $n \to \infty$, which dividing by $\sqrt{n}$, implies

$$a^n \Rightarrow e, \tag{141}$$

as $n \to \infty$.

Proof of Theorem 3.5.2: Since the service discipline is FIFO, the number of customers currently in queue is less than the number that have arrived after the customer currently in service plus one, $A^n(t) - A^n(a^n(t)) + 1$. Additionally, the current queue-length exceeds $A^n(t) - A^n(a^n(t))$ minus the number of customers that have arrived after the one currently in service that will eventually abandon, and so

$$A^n(t) - A^n(a^n(t)) - \sum_{i=A^n(a^n(t))}^{A^n(t)} \mathbf{1}\{V^n(t_i^{n,-}) \geq a_i\} \leq Q^n(t) \leq A^n(t) - A^n(a^n(t)) + 1,$$

or, also using the definition of $\tilde{A}^n$ in (81),

$$\begin{aligned} &\tilde{A}^n(t) - \tilde{A}^n(a^n(t)) + \sqrt{n}\rho^n(t - a^n(t)) - n^{-1/2} \sum_{i=A^n(a^n(t))}^{A^n(t)} \mathbf{1}\{V^n(t_i^{n,-}) \geq a_i^n\} \\ &\leq \tilde{Q}^n(t) \leq \tilde{A}^n(t) - \tilde{A}^n(a^n(t)) + \sqrt{n}\rho^n(t - a^n(t)) + n^{-1/2}. \end{aligned}$$

Subtracting $\tilde{V}^n$ from all sides and adding and subtracting several terms shows

$$\begin{aligned} \left|\tilde{Q}^n(t) - \tilde{V}^n(t)\right| \quad \leq \quad & \left|\tilde{A}^n(t) - \tilde{A}^n(a^n(t))\right| + n^{-1/2} + \left|\tilde{V}^n(t)(\rho^n - 1)\right| \\ & + \left|\rho^n\left(\sqrt{n}(t - a^n(t)) - \tilde{V}^n(a^n(t)^-)\right)\right| + \left|\rho^n\left(\tilde{V}^n(a^n(t)^-) - \tilde{V}^n(t)\right)\right| \\ & + n^{-1/2} \sum_{i=A^n(a^n(t))}^{A^n(t)} \mathbf{1}\left\{V^n(t_i^{n,-}) \geq a_i^n\right\}. \end{aligned} \tag{142}$$

The weak convergence of a^n to the identity process in (141), the functional central limit theorem, and Theorem 3.4.1 imply

$$\tilde{A}^n - \tilde{A}^n \circ a^n \Rightarrow 0 \text{ and } \tilde{V}^n - \tilde{V}^n \circ a^n \Rightarrow 0, \tag{143}$$

as $n \to \infty$. Finally, the inequality (142), (143), the convergence $\rho^n \to 1$ as $n \to \infty$ in (71), the weak convergence in (140), and Lemma 3.5.1 imply the stated result. □

3.5.2 Approximating the Stationary Distribution of the Offered Waiting Time Process

We establish the stationary distributions of the diffusions $\phi^h(W)$ and $\phi^h_C(W)$, and also the average pushing at the upper boundary for the diffusion $\phi^h_C(W)$. We write the stationary distributions in terms of the cumulative hazard rate function $H(x) = \int_0^x h(y)dy$ in order to provide intuition on the condition that the diffusion $\phi^h(W)$ has a unique stationary distribution. Specifically, observe in condition (i) in Proposition 3.5.3 below that if $\theta \leq 0$, then a unique stationary distribution exists because $\phi^h(W)$ is a negative drift diffusion with reflection at the origin. Also, if $\theta > 0$ and there exists z_0 such that $H(z) > \theta$ for all $z > z_0$, then again a unique stationary distribution exists and $\phi^h(W)$ will drift towards $z^\star \equiv \{z : H(z) = \theta\}$ similar to the conventional Ornstein-Uhlenbeck process. Otherwise, if neither of the aforementioned conditions is satisfied, $\phi^h(W)$ has a positive drift and so a stationary distribution does not exist.

Proposition 3.5.3. *Let W be a Brownian motion with drift θ, variance σ^2.*

(i) Suppose there exists z_0 such that $H(z) > \theta$ for all $z > z_0$. Then, the one-sided regulated diffusion $\phi^h(W)$ has a unique stationary distribution π with density

$$p(x) = M \exp\left(\frac{2}{\sigma^2}\left(\theta x - \int_0^x H(s)ds\right)\right),\ x \geq 0,$$

where M is such that $\int_0^\infty p(x)dx = 1$.

(ii) The two-sided regulated diffusion $\phi^h_C(W)$ has a unique stationary distribution π_C with density

$$p_C(x) = M_C \exp\left(\frac{2}{\sigma^2}\left(\theta x - \int_0^x H(s)ds\right)\right),\ 0 \leq x \leq C,$$

and average pushing at the upper boundary

$$\lim_{t\to\infty} t^{-1} E\left[\psi^h_{2,C}(W)(t)\right] = \frac{\sigma^2}{2} \frac{\exp\left(-\int_0^C \frac{2}{\sigma^2}(-\theta + H(x))dx\right)}{\int_0^C \exp\left(-\int_0^y \frac{2}{\sigma^2}(-\theta + H(x))dx\right) dy},$$

where M_C is such that $\int_0^C p(x)dx = 1$.

Furthermore, for any $x \in \Re$, as $t \to \infty$,

$$P\left(\phi^h(W)(t) \leq x\right) \to \pi(x) \text{ and } P\left(\phi^h_C(W)(t) \leq x\right) \to \pi_C(x). \tag{144}$$

The proof of Proposition 3.5.3 requires the following Lemma to establish (144).

Lemma 3.5.4. *Let W be a Brownian Motion with drift θ and variance σ^2. Suppose $\lim_{z\to\infty} H(z) > \theta$. Let*

$$\begin{aligned} T_0 &\equiv \inf\{t \geq 0 : \phi^h(W)(t) = 0\} \\ T_0^C &\equiv \inf\{t \geq 0 : \phi_C^h(W)(t) = 0\} \end{aligned}$$

Then,

$$\begin{aligned} P\left(T_0 < \infty | W(0) = x\right) &= 1, \ \ x \geq 0 \\ P_x\left(T_0^C < \infty | W(0) = x\right) &= 1, \ \ 0 \leq x \leq C. \end{aligned}$$

Proof of Proposition 3.5.3: Echeverria [11] shows that a stationary distribution π of $\phi^h(W)$ ought to satisfy

$$\int_0^\infty (Af)(y)\pi(dy) = 0 \tag{145}$$

for all bounded f that are twice continuously differentiable on $[0, \infty)$ and satisfy $f'(0) = 0$, where

$$Af(y) \equiv \left(\theta - \int_0^y h(u)du\right) f'(y) + \frac{\sigma^2}{2} f''(y).$$

Similarly, a stationary distribution π_C of $\phi_C^h(W)$ ought to satisfy

$$\int_0^C (Af)(y)\pi_C(dy) = 0 \tag{146}$$

for all bounded f that are twice continuously differentiable on $[0, C]$ and satisfy $f'(0) = f'(C) = 0$. It is straightforward to verify (using integration by parts) that π and π_C satisfy (145) and (146) respectively for the desired f. Identical arguments as in the proof of Proposition 1 in [51] then establish that π and π_C are stationary distributions of $\phi^h(W)$ and $\phi_C^h(W)$ respectively.

The average pushing at the upper boundary, $\lim_{t\to\infty} t^{-1} E\left[\psi^h(W)(t)\right]$, follows by arguments mimicking those used to prove Propositions 8 and 9 in [2].

Finally, (144) and the uniqueness of the stationary distribution follow as in Proposition 1 of [51], because for any $x \geq 0$ ($0 \leq x \leq C$), the probability the diffusion $\phi^h(W)$ ($\phi_C^h(W)$) hits 0 in finite time is equal to one by Lemma 3.5.4. □

Table 2: A comparison of the simulated mean queue-length for a GI/GI/1-GI queue with Poisson arrivals at rate 100 per unit, deterministic service with mean 1/100, and abandonment times distributed as given in Column 1.

	E[queue-length]		
Reneging Distribution	Simulated	Approximated	% Error
Deterministic(1)	51.691	50	3.38%
$G(5)$	20.5980	19.8141	3.81%
$G(2)$	11.2930	10.6410	5.77%
$G(1)$=Exponential(1)	6.2585	5.6419	9.85%
$G(0.5)$	3.2545	2.7749	14.74%
$G(0.2)$	1.5029	1.14888	23.56%

Table 3: A comparison of the abandonment probability for a GI/GI/1-GI queue with Poisson arrivals at rate 100 per unit, deterministic service with mean 1/100, and abandonment times distributed as given in Column 1.

	P[abandon]		
Reneging Distribution	Simulated	Approximated	% Error
Deterministic(1)	0.004802	0.005	4.12%
$G(5)$	0.013138	0.013324	1.41%
$G(2)$	0.025864	0.026900	4.01%
$G(1)$=Exponential(1)	0.052728	0.056419	7.00%
$G(0.5)$	0.11273	0.13005	15.36%
$G(0.2)$	0.24128	0.360384	49.36%

3.5.3 Evaluation of the Proposed Diffusion Approximations via Simulation

We begin with a simulation study that explores the effect of variability in the abandonment distribution. As in the introduction, let $G(p)$ be the distribution function associated with a mean 1 gamma random variable having scale and shape parameter p. Observe that such gamma distributions are ordered in variability by the parameter p since $\mathrm{Var}(G(p)) = \frac{1}{p}$ increases as p decreases. The variance of a deterministic distribution is 0, and so the results presented in Tables 2 and 3 are ordered according to the variability of the abandonment distribution.

Each simulation run presented in Tables 2 and 3 assumes Poisson arrivals having rate 100, deterministic service with mean 0.01, and is run to 50,000 time units so has approximately 5,000,000 arrivals. The abandonment distribution varies according to the first column. Recall the drift in our suggested diffusion from (7) for the case that abandonment

Table 4: A comparison of the simulated mean queue-length for a GI/GI/1-GI queue with Poisson arrivals at rate n per unit, deterministic service with mean $1/n$, and abandonment times distributed $G(0.2)$.

	E[queue-length]		
n	Simulated	Approximated	% Error
1000	2.1454	1.793910	16.38%
10,000	3.1025	2.73777	11.76%
100,000	4.4736	4.121112	7.88%
1,000,000	6.5393	6.15025	5.95%
10,000,000	9.7142	9.127725	6.04%
100,000,000	13.9040	13.4975	2.92%

times have a gamma distribution. In the case that abandonment times are deterministic, $F^n(x) = \mathbf{1}\{x \geq C/\sqrt{n}\}$, and so (noting the relationship $H(x) = -\ln(1 - G(x))$ between a distribution function G and its associated cumulative hazard function H)

$$H^n(x) = -\ln\left(1 - F^n(x)\right) = 0, \text{ for } x < \frac{C}{\sqrt{n}},$$

which from (6) implies our suggested approximating diffusion has $H_D^n(x) = 0$ for $x < C$. In both cases, the θ appearing in parts (i) and (ii) of Proposition 3.5.3 is 0 and the variance σ^2 is 1. Note that the stationary density in part (ii) of Proposition 3.5.3 reduces to the uniform distribution on $[0, C]$, which not surprisingly coincides with the limiting result in Theorem 2.1 and Remark 2.2 in Whitt [55] for a standard GI/GI/1 queue with finite waiting room. (Intuitively, from Little's Law, a GI/GI/1 queue with deterministic abandonment times a should resemble a GI/GI/1/λa queue.)

We calculate the approximated steady-state queue-length using Proposition 3.5.3. When $\rho^n \uparrow 1$ as $n \to \infty$, the validity of the desired limit interchange follows from the results of Kingman [25] since the queue-length process in a conventional $GI/GI/1$ queue dominates that in a $GI/GI/1{+}GI$ queue with identical arrival and service processes. Otherwise, when $\rho^n \downarrow 1$ as $n \to \infty$, we assume the validity of the desired limit interchange.

To approximate the probability a customer abandons the system, first observe that $F^n\left(V^n(t_i^{n,-})\right)$ is the probability the ith customer abandons, given the offered waiting time at his arrival. Recalling the definitions of ϵ^n and $\tilde{\epsilon^n}$ in (92) and (112), and the weak

Table 5: A comparison of the simulated abandonment probability for a GI/GI/1-GI queue with Poisson arrivals at rate n per unit, deterministic service with mean $1/n$, and abandonment times distributed $G(0.2)$.

	P[abandon]		
n	Simulated	Approximated	% Error
1000	0.17715	0.233068	31.57%
10,000	0.12769	0.153616	20.30%
100,000	0.090415	0.102426	13.28%
1,000,000	0.063077	0.0687956	9.07%
10,000,000	0.043817	0.046426	5.95%
100,000,000	0.030456	0.0314284	3.19%

convergence $\tilde{\epsilon}^n \Rightarrow 0$ as $n \to \infty$ proved in Subsection 3.4.1.2, under Assumption 1, we find

$$\begin{aligned}\sqrt{n}\frac{\int_0^t F^n\left(V^n(s^-)\right) dA^n(s)}{A^n(t)} &= \frac{n}{A^n(t)}\left(-\tilde{\epsilon}^n(t) + \int_0^t \left(\int_0^{\tilde{V}^n(s^-)} h(w)dw\right) ds\right) \\ &\Rightarrow t^{-1}\int_0^t \left(\int_0^{\phi^h(W)(s)} h(w)dw\right) ds,\end{aligned}$$

as $n \to \infty$, by part (i) of Theorem 3.4.1 and the continuous mapping theorem. Assuming the interchange of limit and expectation, we find that as $n \to \infty$,

$$\begin{aligned}\sqrt{n}E_\pi\left[\frac{\int_0^t F^n\left(V^n(s^-)\right) dA^n(s)}{A^n(t)}\right] &\to t^{-1}\int_0^t E_\pi\left[\int_0^{\phi^h(W)(s)} h(w)dw\right] ds \qquad (147)\\ &= \int_0^\infty \left(\int_0^x h(w)dw\right) p(x)dx,\end{aligned}$$

when the system operates in steady-state, where p is as given in part (i) of Proposition 3.5.3. Similarly, under Assumption 2, recalling the definitions of ϵ_B^n and $\tilde{\epsilon}_B^n$ in (95) and (133), and the weak convergence $\tilde{\epsilon}_B^n \Rightarrow 0$ as $n \to \infty$ argued in the proof of part (ii) of Theorem 3.4.1,

$$\begin{aligned}\sqrt{n}\frac{\int_0^t F^n\left(V^n(s^-)\right) dA^n(s)}{A^n(t)} &= \frac{n}{A^n(t)}\left(-\tilde{\epsilon}_B^n(t) + \int_0^t \left(\int_0^{\tilde{V}^n(s^-)\wedge C} h(w)dw\right) ds + \tilde{U}^n(t)\right) \\ &\Rightarrow t^{-1}\left(\int_0^t \left(\int_0^{\phi_C^h(W)(s)} h(w)dw\right) ds + \psi_{2,C}^h(W)(t)\right),\end{aligned}$$

as $n \to \infty$. Then, by part (ii) of Theorem 3.4.1 and the continuous mapping theorem, assuming the interchange of limit and expectation, we find that, as $n \to \infty$,

$$\begin{aligned}\sqrt{n}E_{\pi_C}\left[\frac{\int_0^t F^n\left(V^n(s^-)\right) dA^n(s)}{A^n(t)}\right] &\to \int_0^\infty \left(\int_0^x h(w)dw\right) p_C(x)dx \qquad (148)\\ &\quad + \frac{\sigma^2}{2}\frac{\exp\left(-\int_0^C \frac{2}{\sigma^2}\left(-\theta + H(x)\right) dx\right)}{\int_0^C \exp\left(-\int_0^y \frac{2}{\sigma^2}\left(-\theta + H(x)\right) dx\right) dy},\end{aligned}$$

where p_C is as given in part (ii) of Proposition 3.5.3. We use the formulas in (147) and (148) to approximate the probability an arriving customer will abandon the system.

Observe the loss in approximation accuracy as the variability of the abandonment distribution increases. This is consistent with our theoretical results which, from Lemmas 3.4.2 and 3.4.6, suggest that our approximations perform better as the fraction of abandoning customers decreases. Increasing the variability of the abandonment distribution increases the recorded abandonment probability in Tables 2 and 3.

However, even highly variable abandonment distributions, such as $G(0.2)$, lead to accurate approximations for fast enough arrival and service rates, and correspondingly small abandonment rates. Tables 4 and 5 consider the worst performing cases in Tables 2 and 3, the $G(0.2)$ case, and shows that the accuracy in our approximations increases as the abandonment probability becomes small. Specifically, we simulate a single-server queue with abandonments having Poisson arrivals at rate n, deterministic service times $1/n$, and customer abandonment times that follow a $G(0.2)$ distribution. We run each simulation to time $n^{-1}5,000,000$ so that approximately $5,000,000$ arrivals occur. Observe that for $n \geq 100,000$, the error in our expected queue-length approximation is under 10%, and for $n \geq 1,000,000$, the error in our approximation for the probability a customer abandons is under 10%.

CHAPTER IV

CONCLUSIONS

In this thesis, we have studied call centers from two different perspectives. In the first half of this thesis, we considered large scale call centers with general service time distributions. Our main results here were to prove convergence results for both the fluid and diffusion scaled queue length process of the $G/GI/N$ queue in the Halfin-Whitt regime. In the second half of this thesis, we considered the phenomena of customer abandonment which may be universally observed in practically all call centers. Our main result in this half of the thesis was to prove convergence results for both the diffusion scaled queue length and workload processes of the $GI/GI/1+GI$ queue in a novel heavy traffic regime which was also introduced for the first time in this thesis. There are many directions which remain for future work on both of these problems.

The $G/GI/N$ queue in the Halfin-Whitt regime has just begun to be studied and there are several problems related to it which at the present moment remain unresolved. For instance, in the future, it would be nice to have a better understand of the limiting process we have obtained in Theorem 5 and Corollary 2. Ideally, one would like to solve for the limiting distribution of this process but unfortunately this in general appears to be a difficult problem. If analytical solutions cannot be found, efficient numerical procedures might perhaps then be developed. Simulation studies could also be conducted to test the accuracy of the proposed approximations based off of the limiting process relative to their actual values. This would be especially interesting when the $G/GI/N$ queue is close to being in the Halfin-Whitt regime.

Customer abandonment is also an interesting topic which recently, due in part to its connection to call centers, has begun to receive a considerable amount of attention, see for instance [58],[56] and [59]. The results we have obtained in this thesis concerning customer abandonment are valid for only a single server queue. However, when modeling call centers,

often multiserver queues are used. It would therefore be interesting to extend the scaling of the hazard rate function introduced in Section 3.2.1 of this thesis to the $G/GI/N+GI$ queue in the Halfin-Whitt regime. One would ideally then like to obtain a corresponding limit for the fluid and diffusion scaled queue length process in this regime. A second interesting topic would also be to use our heavy traffic results to analyze the relationship between capacity and abandonment. It is well known that customer abandonment results in a need for reduced capacity but a further investigation into the specifics of this relationship could potentially yield fruitful insights.

APPENDIX A

REGULATOR MAP PROOFS

In this appendix, we provide the proofs of Proposition 2.2.1 and Lemmas 3.3.2 and 3.3.5. We begin with the proof of Proposition 2.2.1.

A.1 Proof of Proposition 2.2.1

Proof of Proposition 2.2.1.

Suppose first that B is concentrated on the point $c > 0$. In this case it is clear that the solution to (19) satisfies the recursion

$$z(t) = x(t), \quad 0 \leq t < c, \tag{149}$$

and

$$z(t) = x(t) + z^{+}(t - c), \quad t \geq c,$$

in which case it is clearly unique. Furthermore, defining $\varphi_B : D[0, \infty) \mapsto D[0, \infty)$ to be the solution to this recursion, it follows that

$$||\varphi_B(x_1) - \varphi_B(x_2)||_t = ||x_1 - x_2||_t$$

for $0 \leq t < c$. Now suppose that for some integer k, we have

$$||\varphi_B(x_1) - \varphi_B(x_2)||_t \leq k||x_1 - x_2||_t \tag{150}$$

for $(k-1)c \leq t < kc$. It then follows that for $k < t \leq (k+1)c$,

$$\begin{aligned} ||\varphi_B(x_1) - \varphi_B(x_2)||_t &\leq ||x_1 - x_2||_t + ||\varphi_B(x_1) - \varphi_B(x_2)||_{t-c} \\ &\leq ||x_1 - x_2||_t + || + k||x_1 - x_2||_t \\ &= (k+1)||x_1 - x_2||_t. \end{aligned}$$

By induction, this implies that the relationship (150) must hold for all t, which show that φ_B is Lipschitz continuous if B is concentrated on a single point. The proof of measurability

of φ_B for the case of B concentrated at a single point will be included below.

Suppose now that there exists a $\delta > 0$ such that $B(x+\delta) - B(x) < \varepsilon$ for some $0 < \varepsilon < 1$ for all $x \geq 0$. Such a δ will always exist so long as B is not concentrated on a single point. We now provide proofs of existence, uniqueness and Lipschitz continuity for this case

.

Existence: We use the method of successive approximations. Let $u_0 = 0$ and recursively define

$$u_{n+1}(t) = x(t) + \int_0^t u_n^+(s) dB(t-s), \quad t \geq 0.$$

Now observe that

$$u_{n+1}(t) - u_n(t) = \int_0^t (u_n^+(s) - u_{n-1}^+(s)) dB(t-s), \quad t \geq 0,$$

from which it follows that,

$$||u_{n+1} - u_n||_\delta \leq B(\delta)||u_n - u_{n-1}||_\delta < \varepsilon ||u_n - u_{n-1}||_\delta,$$

and so we have the relationship

$$||u_{n+1} - u_n||_\delta \leq \varepsilon^n ||x||_T.$$

Now suppose that for a given integer k,

$$||u_{n+1} - u_n||_{j\delta} \quad \leq \quad n^j \varepsilon^n ||x||_T, \quad \text{for} \quad j = 1, ..., k,$$

and for $n = 1, 2, ...$, so that

$$\begin{aligned} ||u_{n+1} - u_n||_{(k+1)\delta} &\leq \sum_{j=1}^{k+1} \varepsilon ||u_n - u_{n-1}||_{j\delta} + \varepsilon ||u_n - u_{n-1}||_{(k+1)\delta} \\ &\leq kn^k \varepsilon^n ||x||_T + \varepsilon ||u_n - u_{n-1}||_{(k+1)\delta}. \end{aligned} \tag{151}$$

Clearly, we have the relationship

$$||u_2 - u_1||_{(k+1)\delta} \leq (k+1)\varepsilon ||x||_T,$$

and so repeatedly iterating (151) shows that for all $n = 1, 2, ...,$

$$||u_{n+1} - u_n||_{(k+1)\delta} \quad \leq \quad (kn^{k+1} + 1)\varepsilon^n ||x||_T.$$

Therefore, by induction, it follows that the above holds for all k and in particular for $k = \lceil \delta^{-1} T \rceil$. This then implies that

$$||u_{n+1} - u_n||_T \leq ||u_{n+1} - u_n||_{k\delta} \leq kn^k \varepsilon^n ||x||_T \to 0,$$

as $n \to \infty$, which completes the proof for the case of nondegenerate distributions.

Uniqueness: Suppose that u and v both satisfy (19) and let

$$\Delta(t) = u(t) - v(t) = \int_0^t (u(s) - v(s)) dB(t - s), \quad t \geq 0.$$

We then have for $0 \leq t \leq \delta$, that

$$|\Delta(t)| \leq \int_0^t |u(s) - v(s)| dB(t - s) \leq \varepsilon ||\Delta||_\delta,$$

which implies that $\Delta(t) = 0$ on $[0, \delta]$. Next, for $\delta < t \leq 2\delta$, we have that

$$|\Delta(t)| \leq \varepsilon ||\Delta||_\delta + \varepsilon ||\Delta||_{2\delta} = \varepsilon ||\Delta||_{2\delta},$$

which implies that $\Delta(t) = 0$ on $(\delta, 2\delta]$. Iterating the above argument until we reach T completes the proof.

Lipschitz continuity: Note that for $0 \leq t < \delta$, we have

$$||\varphi_B(x_2) - \varphi_B(x_1)||_\delta \leq ||x_2 - x_1||_\delta + \varepsilon ||\varphi_B(x_2) - \varphi_B(x_1)||_\delta,$$

which implies that

$$||\varphi_B(x_2) - \varphi_B(x_1)||_\delta \leq (1 - \varepsilon)^{-1} ||x_2 - x_1||_\delta.$$

Next, for $\delta < t \leq 2\delta$, we have

$$\begin{aligned} ||\varphi_B(x_2) - \varphi_B(x_1)||_{2\delta} \quad &\leq \quad ||x_2 - x_1||_{2\delta} \\ &\quad +\varepsilon ||\varphi_B(x_2) - \varphi_B(x_1)||_\delta + \varepsilon ||\varphi_B(x_2) - \varphi_B(x_1)||_{2\delta} \\ &\leq \quad \frac{1}{(1 - \varepsilon)^2} ||x_1 - x_2||_{2\delta} + \varepsilon ||\varphi_B(x_2) - \varphi_B(x_1)||_{2\delta}, \end{aligned}$$

which implies that

$$||\varphi_B(x_2) - \varphi_B(x_1)||_{2\delta} \quad \leq \quad \frac{1}{(1-\varepsilon)^3}||x_1 - x_2||_{2\delta}.$$

Iterating the above argument for $k = \lceil \delta^{-1}T \rceil - 2$ more time intervals completes the proof.

Finally, we provide a proof of measurability of φ_B for the case of a general B.

Measurability: We begin by defining the function $\Psi_B : D[0,\infty) \to D : [0,\infty)$ by

$$\Psi_B(u)(t) = \int_0^t u(t-s)dB(s), \quad t \geq 0.$$

We will now show that Ψ_B is measurable with respect to the Borel $\sigma-$field $\mathcal{D}$ generated by the Skorohod J_1 topology. Note that since $\mathcal{D}$ is equal to the Kolmogorov σ-field, which is generated by the finite dimensional cylinder sets, it is sufficient to check that for each $n \geq 1$ and $A_1, A_2, ..., A_n \in B(\mathbb{R})$,

$$\{u \in D[0,\infty) : (\Psi_B(u)(t_1), ..., \Psi_B(u)(t_n)) \in (A_1, ..., A_n)\} \in \mathcal{D}$$

for $0 \leq t_1 < t_2, .., < t_n$. However, since σ-algebras are closed under finite intersections, it is sufficient to check that for each $t \geq 0$, $\Psi_B(\cdot)(t)$ is measurable. In order to show this, we will first decompose B into its continuous and discrete parts so that

$$B(t) = B_c(t) + B_d(t), \quad t \geq 0,$$

where we write

$$B_d(t) = \sum_{n=1}^{\infty} c_n \delta_{(p_n)}(t)$$

for the discrete part of B. We will then show that both Ψ_{B_c} and Ψ_{B_d} are measurable functions and so, since the sum of two measurable functions from $(D, \mathcal{D})$ to $(D, \mathcal{D})$ is measurable, and $\Psi_B = \Psi_{B_c} + \Psi_{B_d}$, we will have the desired measurability of Ψ_B.

We begin with the proof of measurability for Ψ_{B_c} for which it will be sufficient to show that for each $t \geq 0$, $\Psi_{B_c}(\cdot)(t)$ is continuous when viewed as a function from $(D[0,\infty), d)$ to $\mathbb{R}$ where d is the Skorohod metric. Let $u_n \to u$ under the metric d. This then implies that

$u_n(t) \to u(t)$ for all but a countable number of t [5]. Furthermore, the measure defined by B_c assigns measure 0 to all countable sets. Thus, since for each $t \geq 0$,

$$\sup_{0 \leq s \leq t} |u_n(s)| \to \sup_{0 \leq s \leq t} |u(s)|,$$

it follows by bounded convergence [12], that

$$\begin{aligned} |\Psi_{B_c}(u_n)(t) - \Psi_{B_c}(u)(t)| &= \left| \int_0^t (u_n(s) - u(s)) dB(t-s) \right| \qquad (152) \\ &\leq \int_0^t |u_n(s) - u(s)| dB(t-s) \\ &\to 0 \text{ as } n \to \infty. \qquad (153) \end{aligned}$$

This completes the proof of the measurability of Ψ_{B_c}.

Now consider Ψ_{B_d}. It is clear that

$$\Psi_{B_d}(u)(t) = \sum_{k=1}^{\infty} \Upsilon_k(u)(t), \quad t \geq 0,$$

where

$$\Upsilon_k(u)(t) = c_k 1\{t \geq p_k\} u(t - p_k).$$

Define

$$\Psi^n_{B_d}(u)(t) = \sum_{k=1}^{n} \Upsilon_k(u)(t), \quad t \geq 0.$$

We have that for each $u \in D[0, \infty)$ and $t \geq 0$,

$$\begin{aligned} \sup_{0 \leq s \leq t} |\Psi_{B^n_d}(u)(s) - \Psi_{B_d}(u)(s)| &= \sup_{0 \leq s \leq t} \left| \sum_{k=n}^{\infty} c_k 1\{s \geq p_k\} u(s - p_k) \right| \\ &\leq \sup_{0 \leq s \leq t} |u(s)| \sum_{k=n}^{\infty} c_k \\ &\to 0 \text{ as } n \to \infty, \end{aligned}$$

and so it follows that $\Psi_{B_d}(u)$ is the pointwise limit of $\Psi_{B^n_d}(u)$ as $n \to \infty$. Thus, if each $\Psi_{B^n_d}(u)$ is measurable, it will follow that Ψ_{B_d} is measurable as well. However, in order to show that $\Psi^n_{B_d}$ is measurable, it will suffice to show that each Υ_k is measurable since the sum of measurable functions is measurable. The fact that Υ_k is measurable may been seen

by noting that Υ_k is first the translation of the function u by a constant p_k and then a multiplication by a constant c_k. Both of these functions are easily seen to be measurable functions and so Υ_k, being the composition of two measurable functions, is measurable as well. This completes the proof of the measurability of Ψ_{B_d}.

Now define that map $\Xi_B : D[0,\infty) \mapsto D[0,\infty)$ by

$$\Xi_B(u)(t) = x(t) + \Psi_B(u)(t) \quad t \geq 0.$$

It is clear that Ξ_B is measurable since Ψ_B is measurable. Furthermore, from the existence portion of the arguments above, it follows that for each $x \in D[0,\infty)$,

$$\varphi(x) = \lim_{n\to\infty} \Xi_B^n(0),$$

where $\Xi_B^n(x) = \Xi_B^{n-1}(\Xi_B(x))$ is the n-fold composition of Ξ with itself and the limit is taken with respect to the metric of uniform convergence over bounded intervals. Thus, since the composition of two measurable functions is measurable, it follows that Ξ_B^n is measurable for each n. But this then implies that φ_B, being the pointwise limit of a sequence of measurable functions, is measurable as well, and so the proof is now complete. □

A.2 Proofs of Lemmas 3.3.2 and 3.3.5

We next give a proof of Lemma 3.3.2.

Proof of Lemma 3.3.2:

Part (i), Existence: Let $T > 0$. Because the function $\eta : D([0,\infty), \Re) \to D([0,\infty), \Re^+)$

$$\eta(w)(t) \equiv \int_0^{\phi(w)(t)} h(\zeta)d\zeta$$

is not Lipschitz cotinuous[1], Lemma 1 in Reed and Ward [42] is not directly applicable. However, define w_0 to be the zero process, and

$$w_{n+1}(t) = x(t) - \int_0^t \left(\int_0^{\phi(w_n)(s)} h(u)du \right) ds. \tag{154}$$

Suppose there exists $M > 0$ such that

$$\|w_n\|_T \leq M \text{ for all } n \geq 0. \tag{155}$$

[1]Note the function η would be Lipschitz continuous if h were bounded on $[0,\infty)$.

Then, since the definition of ϕ in (96) implies

$$\|\phi(w_n)\|_T \leq 2\|w_n\|_T \leq 2M,$$

we find that for any $0 \leq s \leq T$,

$$\int_0^{\phi(w_n)(s)} h(\zeta)d\zeta = \int_0^{\phi(w_n)(s)} (h(\zeta) \wedge \|h\|_{2M})\, d\zeta,$$

and so

$$\begin{aligned} \|\eta(w_{n+1}) - \eta(w_n)\|_T &= \sup_{0\leq t\leq T} \int_{\phi(w_n)(t)\wedge\phi(w_{n+1})(t)}^{\phi(w_n)(t)\vee\phi(w_{n+1}(t))} (h(\zeta) \wedge \|h\|_{2M})\, d\zeta \qquad (156) \\ &\leq \|h\|_{2M}\|\phi(w_{n+1}) - \phi(w_n)\|_T \\ &\leq 2\|h\|_{2M}\|w_{n+1} - w_n\|_T, \end{aligned}$$

where the last inequality follows from the Lipschitz continuity of ϕ noted in (102). The inequality (156) implies the arguments used to prove existence in Lemma 1 of Reed and Ward [42] are valid when the constant κ in their proof is taken to be $2\|h\|_{2M}$.

We now show (155) to complete the proof. Since the definition of ϕ in (96) implies $\phi(w_0)(0) = 0$, $w_1 = x$. Next, from (154), because h is assumed non-negative and ϕ defined in (96) is also positive,

$$w_n \leq w_1 \text{ for all } n \geq 1. \qquad (157)$$

Lemma 5.1 in Kruk et al [29] establishes that for all $n \geq 2$

$$\phi(w_n) \leq \phi(w_1). \qquad (158)$$

To see the conditions of Lemma 5.1 in [29] are satisfied, observe that

$$w_n(t) = w_1(t) - \int_0^t \left(\int_0^{\phi(w_{n-1})(s)} h(u)du \right) ds$$

is written as the difference of w_1 and a non-decreasing function. Use of (158) shows that for all $n \geq 3$

$$\begin{aligned} w_n(t) &= x(t) - \int_0^t \left(\int_0^{\phi(w_{n-1})(s)} h(u)du \right) ds \qquad (159) \\ &\geq x(t) - \int_0^t \left(\int_0^{\phi(w_1)(s)} h(u)du \right) ds = w_2(t), \end{aligned}$$

for all $t \geq 0$, since h is assumed non-negative. Combining (157) and (159), we conclude

$$w_2 \leq w_n \leq w_1 \text{ for all } n \geq 2. \tag{160}$$

Set $M = \|w_1\|_T \vee \|w_2\|_T$. Then, (160) implies (155) is valid.

Part (i), Uniqueness: Suppose both u and v satisfy (100). As in (158) in the proof of existence, Lemma 5.1 in Kruk et al [29] establishes

$$\|\phi(u)\|_T \leq \|\phi(x)\|_T \text{ and } \|\phi(v)\|_T \leq \|\phi(x)\|_T. \tag{161}$$

Set $N \equiv \|\phi(x)\|_T$. Use of the integral equation definition in (100) and (161) shows

$$\begin{aligned} \triangle(t) \equiv u(t) - v(t) &= \int_0^t \left(\int_0^{\phi(v)(s)} h(x)dx - \int_0^{\phi(u)(s)} h(x)dx \right) ds \\ &= \int_0^t \left(\int_0^{\phi(v)(s)} (h(\zeta) \wedge ||h||_N)\, d\zeta - \int_0^{\phi(u)(s)} (h(\zeta) \wedge ||h||_N)\, d\zeta \right) ds, \end{aligned}$$

and so the Lipschitz continuity of ϕ in (102) implies

$$\begin{aligned} |\triangle(t)| &\leq \int_0^t \left| \int_0^{\phi(v)(s)} (h(\zeta) \wedge ||h||_N)\, d\zeta - \int_0^{\phi(u)(s)} (h(\zeta) \wedge ||h||_N)\, d\zeta \right| ds \\ &\leq \int_0^t ||h||_N \, |\phi(v)(s) - \phi(u)(s)|\, ds \\ &\leq 2t||h||_N \|\triangle\|_t, \end{aligned}$$

which implies $\triangle(t) = 0$ for all $0 \leq t \leq (2||h||_N)^{-1}$. For $(2||h||_N)^{-1} < t < 2\,(2||h||_N)^{-1}$,

$$\begin{aligned} |\triangle(t)| &\leq \|\triangle\|_{(2||h||_N)^{-1}} + \left(t - (2||h||_N)^{-1} \right) 2||h||_N \|\triangle\|_{2(2||h||_N)^{-1}} \\ &\leq \|\triangle\|_{2(2||h||_N)^{-1}}, \end{aligned}$$

and so $\triangle(t) = 0$ for all $0 \leq t \leq 2\,(2||h||_N)^{-1}$. Continued iteration of this argument implies $\triangle = \vec{0}$.

Part (ii): From the Lipschitz continuity of ϕ noted in (102) and assumption,

$$\begin{aligned} \|\phi\,(x_j)\,\|_T &\leq \|\phi(x)\|_T + \|\phi(x_j) - \phi(x)\|_T \\ &\leq \|\phi(x)\|_T + 2, \quad j \in \{1, 2\}. \end{aligned} \tag{162}$$

As in (158) in the proof of existence in part (i), Lemma 5.1 in Kruk et al [29] establishes

$$\|\phi\left(\mathcal{M}^h(x_j)\right)\|_T \leq \|\phi(x_j)\|_T,\ j \in \{1,2\},$$

and so from (162),

$$\|\phi\left(\mathcal{M}^h(x_1)\right)\|_T \vee \|\phi\left(\mathcal{M}^h(x_1)\right)\|_T \leq \|\phi(x)\|_T + 2. \tag{163}$$

Set $\overline{c} \equiv \|\phi(x)\|_T + 2$, and observe from the definition of the mapping $\mathcal{M}^h$ in (100), the inequality (163), and the Lipschitz continuity of ϕ noted in (102), that for $0 \leq t \leq T$,

$$\begin{aligned}
\|\mathcal{M}^h(x_1) - \mathcal{M}^h(x_2)\|_t &\leq \|x_1 - x_2\|_t + \sup_{0\leq s\leq t}\left|\int_0^s \left(\int_{\phi(\mathcal{M}^h(x_1))(u)}^{\phi(\mathcal{M}^h(x_2))(u)} h(\zeta)d\zeta\right) du\right| \\
&\leq \|x_1 - x_2\|_t + t\|h\|_{\overline{c}}\|\phi\left(\mathcal{M}^h\left(x_2\right)\right) - \phi\left(\mathcal{M}^h\left(x_1\right)\right)\|_t \\
&\leq \|x_1 - x_2\|_t + 2t\|h\|_{\overline{c}}\|\mathcal{M}^h\left(x_2\right) - \mathcal{M}^h\left(x_1\right)\|_t.
\end{aligned}$$

Therefore, for any $0 \leq t \leq (4\|h\|_{\overline{c}})^{-1}$,

$$\|\mathcal{M}^h(x_2) - \mathcal{M}^h(x_1)\|_t \leq \frac{\|x_1 - x_2\|_t}{1 - 2\|h\|_{\overline{c}}t}. \tag{164}$$

For $(4\|h\|_{\overline{c}})^{-1} < t \leq (2\|h\|_{\overline{c}})^{-1}$,

$$\begin{aligned}
\|\mathcal{M}^h(x_2) - \mathcal{M}^h(x_1)\|_t &\leq \|\mathcal{M}^h(x_2) - \mathcal{M}^h(x_1)\|_{\frac{1}{4\|h\|_{\overline{c}}}} + 2\|x_1 - x_2\|_t \\
&\quad + \sup_{0\leq s\leq t}\left|\int_{\frac{1}{4\|h\|_{\overline{c}}}}^s \left(\int_{\phi(\mathcal{M}^h(x_1))(u)}^{\phi(\mathcal{M}^h(x_2))(u)} h(\zeta)d\zeta\right) du\right|,
\end{aligned}$$

and so, also using (164), the inequality (163), and (102),

$$\|\mathcal{M}^h(x_2) - \mathcal{M}^h(x_1)\|_t \leq 4\|x_1 - x_2\|_t + 2\left(t - \frac{1}{4\|h\|_{\overline{c}}}\right)\|h\|_{\overline{c}}\|\mathcal{M}^h(x_2) - \mathcal{M}^h(x_1)\|_t,$$

or

$$\|\mathcal{M}^h(x_2) - \mathcal{M}^h(x_1)\|_t \leq \frac{4\|x_1 - x_2\|_t}{1 - 2\|h\|_{\overline{c}}\left(t - \frac{1}{4\|h\|_{\overline{c}}}\right)}.$$

Since only a finite number of intervals of length $(4\|h\|_{\overline{c}})^{-1}$ partition the interval $[0,T]$, continued iteration of the above argument establishes

$$\|\mathcal{M}^h(x_2) - \mathcal{M}^h(x_1)\|_t \leq \kappa\|x_1 - x_2\|_t$$

for κ finite (but dependent on x through $\bar{c}$) and any $0 \leq t \leq T$.

Part (iii): Proof. Let $w \in D([0,\infty),\Re)$ be the unique solution to (100) for $x \in D([0,\infty),\Re)$. Note that since $x \in D([0,\infty),\Re)$, it follows that for each $T \geq 0$, there exists an $M \geq 0$ such that $\sup_{0\leq t\leq T}|x(t)| \leq (M/2-1)$. The following observation will be useful. Because $\int_0^t \left(\int_0^{\phi(w)(s)} h(u)du\right) ds$ is non-decreasing in t, Lemma 5.1 in Kruk et al shows $\phi(w) \leq \phi(x)$. Thus, since for any $T \geq 0$, $\phi(x)(t) \leq 2\sup_{0\leq s\leq t}|x(s)| \leq M-2$, the following inequality is valid

$$\phi(w)(t) \leq M-2 \text{ for all } 0 \leq t \leq T. \tag{165}$$

Suppose now that $x^n \to x$ as $n \to \infty$ in the Skorohod J_1 topology and let T be a continuity point of x. Then, there must exist a sequence of absolutely continuous homeomorphisms $\{\lambda^n\}$ of $[0,T]$ such that

$$||x^n \circ \lambda^n - x||_T \vee ||\lambda^n - e||_T \to 0.$$

Furthermore, it suffices to consider absolutely continuous homeomorphisms, such that

$$||x^n \circ \lambda^n - x||_T \vee ||\dot{\lambda}^n - 1||_T \to 0$$

as $n \to \infty$, see Billingsley [4] for more details. Also, for n sufficiently large we have that $\sup_{0\leq t\leq T}|x^n(t)| \leq \sup_{0\leq t\leq T}|x(t)|+1 \leq M/2$ and so reasoning similar to the above implies that

$$\phi(w^n)(t) \leq M \text{ for all } 0 \leq t \leq T,$$

where $w^n = \mathcal{M}^h(x^n)$ is the solution to (100) for x^n.

Now, for all $0 \leq t \leq T$ and n sufficiently large,

$$\begin{aligned}
& ||w^n \circ \lambda^n - w||_t \\
= & \left\| x^n \circ \lambda^n - x - \int_0^{\lambda^n} \left(\int_0^{\phi(w^n)(s)} h(u)du \right) ds - \int_0^e \left(\int_0^{\phi(w)(s)} h(u)du \right) ds \right\|_t \qquad (166) \\
= & \left\| x^n \circ \lambda^n - x - \int_0^e \left(\int_0^{\phi(w^n)(\lambda^n(s))} h(u)du \right) \dot{\lambda}^n(s) ds - \int_0^e \left(\int_0^{\phi(w)(s)} h(u)du \right) ds \right\|_t \\
\leq & ||x^n \circ \lambda^n - x||_t + ||\dot{\lambda}^n(s) - 1||_t \int_0^t \left(\int_0^{\phi(w^n)(\lambda^n(s))} h(u)du \right) ds \\
& + \left\| \int_0^e \left(\int_0^{\phi(w^n)(\lambda^n(s))} h(u)du - \int_0^{\phi(w)(s)} h(u)du \right) ds \right\|_t . \\
\leq & \|x^n \circ \lambda_n - x\|_t + \|\dot{\lambda}^n(s) - 1\|_t \|h\|_M MT \\
& + \|h\|_M \int_0^t \sup_{0 \leq s \leq t} |\phi(w_n)(\lambda^n(s)) - \phi(w)(s)| \, ds.
\end{aligned}$$

By Lemma 13.5.2 in [54] $\phi(w^n)(\lambda^n(s)) = \phi(w^n \circ \lambda^n)(s)$ and by Lemma 13.5.1 in [54], ϕ is Lipschitz continuous with respect to the uniform metric with Lipschitz constant 2, and so

$$\begin{aligned}
& \int_0^t \sup_{0 \leq s \leq t} |\phi(w^n)(\lambda^n(s)) - \phi(w)(s)| \, ds \qquad (167) \\
= & \int_0^t \sup_{0 \leq s \leq t} |\phi(w^n \circ \lambda^n)(s) - \phi(w)(s)| \, ds \\
\leq & \int_0^t 2\|w^n \circ \lambda^n - w\|_s ds.
\end{aligned}$$

We conclude from (166) and (167) that

$$\|w^n \circ \lambda^n - w\|_t \leq \|x^n \circ \lambda^n - x\|_t + \|\dot{\lambda}^n(s) - 1\|_t \|h\|_M MT + 2\|h\|_M \int_0^t \|w^n \circ \lambda^n - w\|_s ds.$$

Let $\varepsilon > 0$ be arbitrarily small. Then, there exists n_0 such that

$$\|x^n \circ \lambda^n - x\|_t + \|\dot{\lambda}^n(s) - 1\|_t \|h\|_M MT \leq \epsilon$$

for $n \geq n_0$. We then have that

$$\|w^n \circ \lambda^n - w\|_t \leq \epsilon + 2\|h\|_M \int_0^t \|w^n \circ \lambda^n - w\|_s ds$$

for $0 \leq t \leq T$ and $n \geq n_0$. Therefore, by Gronwall's inequality,

$$||w^n \circ \lambda^n - w||_T \leq \epsilon e^{2T\|h\|_M}.$$

Since also $|||\dot{\lambda}^n - 1||_T \to 0$ as $n \to \infty$, it follows that

$$\|w^n \circ \lambda^n - w\|_T \vee \|\lambda^n - e\|_T \to 0,$$

as $n \to \infty$. □

Finally, we now have a proof of Lemma 3.3.5.

Proof of Lemma 3.3.5: From Lemma 1 in [42], for parts (i) and (ii) it is sufficient to verify the Lipschitz continuity of the function $\eta : D([0,\infty), \Re) \to D([0,\infty), \Re)$, defined as

$$\eta(w)(t) \equiv \int_0^{\phi_C(w)(t)} h(u)du$$

for $w \in D([0,\infty), \Re)$. Since

$$\|\eta(w_1) - \eta(w_2)\|_T \leq \sup_{0\leq t\leq T} \int_{\phi_C(w_2)(t)}^{\phi_C(w_1)(t)} |h(u)|du \leq \|h\|_C \|\phi_C(w_1) - \phi_C(w_2)\|_T,$$

and Theorem 14.8.1 in Whitt [54] establishes the mapping ϕ_C is Lipschitz continuous with Lipschitz constant 2, we conclude

$$\|\eta(w_1) - \eta(w_2)\|_T \leq 2\|h\|_C \|w_1 - w_2\|_T.$$

Note that if the condition $0 \leq x_1(0), x_2(0) \leq C$ is not satisfied, then the above inequality is not valid, and must also accommodate the jump at time 0.

The proof of part (iii) proceeds in a similar manner to the proof of part (iii) of Lemma 3.3.2 and therefore has not been included. It is, however, necessary to note that using the explicit form of the two-sided regulator mapping in Kruk et al [29],

$$\phi_C(x)(t) = \phi(x)(t) - \sup_{0\leq s\leq t} \left([\phi(x)(s) - C]^+ \wedge \inf_{s\leq u\leq t} \phi(x)(u) \right),$$

it is straightforward to show that

$$\phi_C(x)(\lambda(t)) = \phi_C\left(x \circ \lambda\right)(t).$$

□

APPENDIX B

G/GI/N QUEUE PROOFS

In this appendix, we provide the proofs of Propositions 2.3.2 and 2.4.2. Our proof of Proposition 2.3.2 will closely parallel the proofs of Lemmas 3.4 through 3.8 of [28]. In order to begin, we must first set up the following notation. Let us define the two parameter process

$$V^N(t,x) = \sum_{i=1}^{\hat{A}^N(t)} (1\{\eta_i \leq x\} - F(x)), \quad t \geq 0, \ x \geq 0, \tag{168}$$

where we recall from Section 2.4 that $\hat{A}^N$ is defined to be the number of customers who have begun service by time t and η_i is the service time of the i^{th} customer to arrive to the system after time zero as defined in Section 2.1.2. Note that by setting

$$U^N(t,x) = \sum_{i=1}^{\lfloor Nt \rfloor} (1\{F(\eta_i) \leq x\} - x), \quad t \geq 0, \ 0 \leq x \leq 1,$$

we have

$$V^N(t,x) = U^N(\check{A}^N(t), F(x)), \tag{169}$$

where

$$\check{A}^N(t) = \frac{\hat{A}^N(t)}{N}, \quad t \geq 0. \tag{170}$$

It then follows from the definition of M_2^N in (12) that

$$M_2^N(t) = \int_0^t \int_0^t 1\{s + x \leq t\} dV^N(s,x), \tag{171}$$

where the integrals above are taken over the closed intervals $[0,t]$. We will now decompose M_2^N in two processes, G^N and H^N. Let

$$L^N(t,x) = \sum_{i=1}^{\hat{A}^N(t)} \left(1\{\eta_i \leq x\} - \int_0^{x \wedge \eta_i} \frac{dF(y)}{1 - F(y-)} \right), \quad t \geq 0, \ x \geq 0, \tag{172}$$

where $F(y-) = \lim_{x \to y} F(x)$.

By (168) and (172), we have that

$$V^N(t,x) = -\int_0^x \frac{V^N(t,y-)}{1-F(y-)} dF(y) + L^N(t,x). \tag{173}$$

Therefore, by (171) and (173), we have

$$M_2^N(t) \quad = \quad G^N(t) + H^N(t), \tag{174}$$

where

$$G^N(t) = -\int_0^t \frac{V^N(t-x,x-)}{1-F(x-)} dF(x), \quad t \geq 0,$$

and

$$H^N(t) = \int_0^t \int_0^t 1\{s+x \leq t\} dL^N(s,x), \quad t \geq 0. \tag{175}$$

We set $G^N = \{G^N(t), t \geq 0\}$ and $H^N = \{H^N(t), t \geq 0\}$ and note that (174) is the desired decomposition. It will be useful in proving several results related to M_2^N such as tightness and weak convergence.

Now let

$$H_k^N(t) = \sum_{i=1}^{\hat{A}^N(t) \wedge k} \left(1\{0 < \eta_i \leq t - \hat{\tau}_i^N\} - \int_{0+}^{\eta_i \wedge (t-\hat{\tau}_i^N)^+} \frac{dF(u)}{1-F(u-)} \right),$$

for $t \geq 0$, where

$$\hat{\tau}_i^N = \inf\{t \geq 0 : \hat{A}^N(t) \geq i\}$$

is the time at which the i^{th} customer to enter service after time zero begins being served. We set $H_k^N = \{H_k^N(t), t \geq 0\}$. Furthermore, define the filtration $\mathbf{H}^N = (\mathcal{H}_t^N, t \geq 0)$ by

$$\begin{aligned} \mathcal{H}_t^N \quad = \quad & \sigma\{\xi_i, i = 1, ..., \hat{A}^N(t)\} \\ & \vee \sigma\{1\{\eta_i = 0\}, 1\{\eta_i \leq s - \hat{\tau}_i^N\}, s \leq t, i = 1, ..., \hat{A}^N(t)\} \\ & \vee \sigma\{\hat{A}^N(s), s \leq t\} \vee \mathcal{N}, \end{aligned}$$

where ξ_i is as defined in (8) of Section 2.1.2 and $\mathcal{N}$ is the $\mathbb{P}$ completion of $\mathcal{F}$. It easy to see that $\mathbf{H}^N$ satisfies the usual conditions and is actually a filtration.

The following lemma appears in [28]. The proof in our case is similar to the one there and as a consequence we will only point out the minor differences.

Lemma B.0.1. *The process H_k^N is an $\boldsymbol{H}^N$-square-integrable martingale with predictable quadratic variation process*

$$\langle H_k^N \rangle(t) = \sum_{i=1}^{\hat{A}^N(t)\wedge k} \int_{0+}^{\eta_i \wedge (t-\hat{\tau}_i^N)^+} \frac{1-F(u)}{(1-F(u-))^2} dF(u), \quad t \geq 0.$$

Proof. We only point out the small changes to the proof of Lemma 3.5 of [28] which need to made.

Equation (3.40) on page 257 in the proof of Lemma 3.5 of [28] should be switched to

$$\{\hat{\tau}_{i+1}^N = \hat{\tau}_i^N\} \cap \{\eta_i > 0\} = \{h_{i+1}^N(\xi_r, r \geq 1, \eta_p, p \geq 1, p \neq i) \leq \hat{\tau}_i^N\} \cap \{\eta_i > 0\},$$

where the random variable h_{i+1}^N is the time of the $(i+1)^{st}$ departure from the system assuming that customer i's service time is infinitely long. Equation (3.41) should also be similarly changed by letting h_r^N be the time of the r^{th} departure from the system assuming that customer i's service time is infinitely long.

The process $\tilde{A}^N(u)$ defined in the middle paragraph on page 258 of [28] should now be the number of customers entering service by time u if customer i's service time was infinitely long.

The random variable $\tilde{\tau}_i^N$ on page 260 of [28] should now be set equal to the time at which customer i enters service assuming that customer j's service time is infinitely long.

□

Now note that by (22) and (174), we have

$$\bar{M}_2^N(t) = \bar{G}^N(t) + \bar{H}^N(t), \tag{176}$$

where

$$\bar{G}^N = \frac{G^N}{N} \tag{177}$$

and

$$\bar{H}^N = \frac{H^N}{N}. \tag{178}$$

It therefore follows by (176) that in order to $\bar{M}_2^N \Rightarrow 0$ as $N \to \infty$, it will be sufficient to show that $\bar{G}^N$ and $\bar{H}^N$ each converge to 0 separately. We begin with $\bar{G}^N$.

First, however, let us set

$$\bar{U}^N = \frac{U^N}{N}. \tag{179}$$

We then have the following.

Lemma B.0.2. *$\bar{G}^N \Rightarrow 0$ as $N \to \infty$.*

Proof. We will first show that for each $\delta > 0$ and $T > 0$,

$$\lim_{\varepsilon \downarrow 0} \limsup_N P\left(\sup_{0 \le t \le T} \left| \int_0^t \frac{\bar{V}^N(t-x, x-)}{1-F(x-)} 1\{F(x-) > 1-\varepsilon\} dF(x) \right| > \delta \right) = 0, \tag{180}$$

where $\bar{V}^N(t,x) = N^{-1} V^N(t,x)$. The proof of this will be identical to the proof in [28] but for completeness we will include it here as well.

In view of (169), and recalling the definition of $\bar{U}^N$ from (179), we have for any $k > 0$,

$$\begin{aligned} & P\left(\sup_{0 \le t \le T} \left| \int_0^t \frac{\bar{V}^N(t-x, x-)}{1-F(x-)} 1\{F(x-) > 1-\varepsilon\} dF(x) \right| > \delta \right) \\ \le\ & P(\check{A}^N(T) > kT) \\ & + P\left(\int_0^\infty \frac{1\{F(x-) > 1-\varepsilon\}}{1-F(x-)} \sup_{0 \le t \le kT} |\bar{U}^N(t, F(x-))| dF(x) > \delta \right). \end{aligned}$$

For k sufficiently large, we have by the definition of $\check{A}^N$ in (170) and Lemma 2.4.1 of Section 2.4 that

$$P(\check{A}^N(T) > kT) \to 0 \ \text{ as } \ N \to \infty.$$

Therefore, by applying Chebyshev's inequality and Fubini's theorem, we reduce our task to proving

$$\lim_{\varepsilon \downarrow 0} \limsup_N \int_0^\infty \frac{1\{F(x-) > 1-\varepsilon\}}{1-F(x-)} E \sup_{0 \le t \le kT} |\bar{U}^N(t, F(x-))| dF(x) = 0.$$

As in Lemma 3.1 of [28], for any fixed $x \ge 0, \{\bar{U}^N(t, F(x-)), t \ge 0\}$ is a locally square-integrable martingale with respect to the filtration $\mathbf{G}^N = \mathcal{G}^N(t), t \ge 0$ defined by $\mathcal{G}^N(t) = \{\eta_i, 1 \le i \le \lfloor nt \rfloor\} \vee \mathcal{N}$. Moreover, it has the predictable quadratic-variation process, see [36] and [21],

$$\langle \bar{U}^N(\cdot, F(x-)) \rangle(t) = \frac{\lfloor Nt \rfloor}{N^2} F(x-)(1-F(x-)), \quad t \ge 0.$$

By Theorem 1.9.5 in [36], we have that

$$\begin{aligned} E\left[\sup_{0\le t\le kT}|\bar{U}^N(t,F(x-))|\right] &\le 3E[\langle \bar{U}^N(\cdot,F(x-))\rangle(kT)]^{1/2} \\ &\le 3\sqrt{kT}(F(x-)(1-F(x-)))^{1/2}, \end{aligned}$$

which implies that

$$\begin{aligned} &\int_0^\infty \frac{1\{F(x-)>1-\varepsilon\}}{1-F(x-)} E\left[\sup_{0\le t\le kT}|\bar{U}^N(t,F(x-))|\right] dF(x) \\ &\le 3\sqrt{kT}\int_0^\infty \frac{1\{F(x-)>1-\varepsilon\}}{1-F(x-)} dF(x). \end{aligned}$$

Denoting $F^{-1}(x)=\inf\{y: F(y)>x\}$, we have, by a change of variables, that the latter equals

$$\begin{aligned} 3\sqrt{kT}\int_0^1 \frac{1\{F(F^{-1}(x)-)>1-\varepsilon\}}{(1-F(F^{-1}(x)-))^{1/2}}dx &\le 3\sqrt{kT}\int_0^1 \frac{1\{x>1-\varepsilon\}}{(1-x)^{1/2}}dx \\ &\le 6\sqrt{kT}\sqrt{\varepsilon}\to 0 \ \text{ as } \ \varepsilon\to 0. \end{aligned}$$

The proof of (180) is completed.

Next, first note that by Lemma 3.1 in [28], it follows that

$$\bar{U}^N \Rightarrow 0 \ \text{ as } \ N\to\infty,$$

where $\bar{U}^N=\{\bar{U}^N(t,x), t\ge 0, 0\le x\le 1\}$. This then implies that for each $k\ge 0$,

$$\sup_{0\le t\le kT}\sup_{0\le x\le 1}|\bar{U}^N(t,F(x-))|\Rightarrow 0 \ \text{ as } \ N\to\infty.$$

Hence, for each $\varepsilon>0$ and $\delta>0$, we have

$$\begin{aligned} & P\left(\int_0^\infty \frac{1\{F(x-)\le 1-\varepsilon\}}{1-F(x-)}\sup_{0\le t\le kT}|\bar{U}^N(t,F(x-))|dF(x)>\delta\right) \\ \le\ & P\left((1-\varepsilon)^{-1}\sup_{0\le t\le kT}\sup_{0\le x\le 1}|\bar{U}^N(t,F(x-))|>\delta\right) \\ \to\ & 0 \ \text{ as } \ N\to\infty, \end{aligned}$$

which, when combined with (180), completes the proof. □

We will next show that $\bar{H}^N$ converges to 0 as N goes to ∞. Again, the modifications to the proof of Lemma 3.7 of [28] are slight but we include a full proof for completeness.

Lemma B.0.3. *$\bar{H}^N \Rightarrow 0$ as $N \to \infty$.*

Proof. Let

$$\hat{H}^N(t) = N^{-1} \sum_{i=1}^{\hat{A}^N(t)} \left(1\{0 < \eta_i \leq t - \hat{\tau}_i^N\} - \int_{0+}^{\eta_i \wedge (t-\hat{\tau}_i^N)^+} \frac{dF(u)}{1-F(u-)} \right), \quad t \geq 0.$$

By (175), (172) and (178) we have that

$$\bar{H}^N(t) = N^{-1} \sum_{i=1}^{\hat{A}^N(t)} (1\{\eta_i = 0\} - F(0)) + \hat{H}^N(t).$$

We will first show that the term involving the summation converges to 0. Let $T \geq 0$ and $\delta > 0$. We have

$$\begin{aligned} & P\left(\sup_{0 \leq t \leq T} \left| N^{-1} \sum_{i=1}^{\hat{A}^N(t)} (1\{\eta_i = 0\} - F(0)) \right| > \delta \right) \\ & \leq P(N^{-1}\hat{A}^N(T) > k) + P\left(\sup_{0 \leq t \leq 1} \left| N^{-1} \sum_{i=1}^{\lfloor Nkt \rfloor} (1\{\eta_i = 0\} - F(0)) \right| > \delta \right). \end{aligned}$$

However, for sufficiently large k, we have by the definition of $\hat{A}^N$ and assumption (15) of Section 2.1.2 that

$$P(N^{-1}\hat{A}^N(T) > k) \quad \leq \quad P(\bar{A}^N(T) > k) \to 0 \ \text{ as } \ N \to \infty. \tag{181}$$

Furthermore, by the functional strong law of large numbers and the i.i.d. assumption of $\{\eta_i, i \geq 1\}$, it follows that

$$P\left(\sup_{0 \leq t \leq 1} \left| N^{-1} \sum_{i=1}^{\lfloor Nkt \rfloor} (1\{\eta_i = 0\} - F(0)) \right| > \delta \right) \to 0 \ \text{ as } \ N \to \infty.$$

It thus remains to show the convergence of $\hat{H}^N$ to 0.

Fix $T > 0$. For each $\varepsilon > 0$, we have

$$\begin{aligned} & P\left(\sup_{0 \leq t \leq T} \hat{H}^N(t) > \varepsilon \right) \\ & \leq P(N^{-1}\hat{A}^N(T) > k) + P\left(\sup_{0 \leq t \leq T} |\bar{H}_{Nk}^N(t)| > \varepsilon \right), \end{aligned}$$

where

$$\bar{H}^N_{Nk} = \frac{H^N_{Nk}}{N}.$$

By (181), we have that for k sufficiently large,

$$P(N^{-1}\hat{A}^N(T) > k) \to 0 \text{ as } N \to \infty.$$

Next, by Lemma B.0.1, we have that $\bar{H}^N_{Nk}$ is an $\mathbf{H}^N$-square-integrable martingale with predictable quadratic variation process

$$\langle \bar{H}^N_{Nk} \rangle(t) = N^{-2} \sum_{i=1}^{\hat{A}^N(t)\wedge Nk} \int_{0+}^{\eta_i \wedge (t-\hat{\tau}^N_i)^+} \frac{1-F(u)}{(1-F(u-))^2} dF(u), \quad t \geq 0. \tag{182}$$

Thus, by the Lenglart-Rebolledo inequality [36], for any $\gamma > 0$,

$$P\left(\sup_{0\leq t\leq T} |\bar{H}^N_{Nk}(t)| > \varepsilon \right) \leq \frac{\gamma}{\varepsilon^2} + P(\langle \bar{H}^N_{Nk} \rangle(T) > \gamma).$$

However, by (182),

$$\langle \bar{H}^N_{Nk} \rangle(T) \leq N^{-2} \sum_{i=1}^{A^N(T)} \int_0^{\eta_i} \frac{dF(u)}{1-F(u-)}. \tag{183}$$

Furthermore, since $E[\int_0^{\eta_i}(1-F(u-))^{-1}dF(u)] = 1$, it follows by the functional strong law of large numbers that

$$N^{-2} \sum_{i=1}^{\lfloor N\cdot \rfloor} \int_0^{\eta_i} \frac{dF(u)}{1-F(u-)} \Rightarrow 0 \text{ as } N \to \infty.$$

By (183), the random time change theorem and assumption (15), this then implies that for any $\gamma > 0$,

$$P(\langle \bar{H}^N_{Nk} \rangle(T) > \gamma) \to 0 \text{ as } N \to \infty,$$

which completes the proof. □

We are now in a position to give a proof of Proposition 2.3.2.

Proof of Proposition 2.3.2. The proof follows by the decomposition (176) and Lemmas B.0.2 and B.0.3 above. □

The remainder of the appendix will now be devoted to providing a proof of Proposition 2.4.2. We begin by defining the processes

$$\tilde{G}^N = \frac{G^N}{\sqrt{N}}$$

and

$$\tilde{H}^N = \frac{H^N}{\sqrt{N}},$$

and note that by (41) and (174) it follows that

$$\tilde{M}_2^N = \tilde{G}^N + \tilde{H}^N. \tag{184}$$

Our first result will be to show that the sequence $\{\tilde{M}_2^N\}$ is tight. In order to show this, it will be sufficient to show that both $\{\tilde{G}^N\}$ and $\{\tilde{H}^N\}$ are tight. We begin with a proof for $\{\tilde{G}^N\}$.

Lemma B.0.4. *The sequence $\{\tilde{G}^N\}$ is tight.*

Proof. By virtue of Lemma 2.4.1 and the fact that the identity process $e(t) = t$ is a continuous process, the proof now follows identically to the proof of Lemma 3.4 in [28]. The modifications to this proof are essentially trivial and the interested reader is referred to Lemma 3.4 of [28] for further details. □

Next, we show that $\{\tilde{H}^N\}$ is tight.

Lemma B.0.5. *The sequence $\{\tilde{H}^N\}$ is tight.*

Proof. Since by Lemma B.0.1, the process H_k^N is an $\mathbf{H}^N$-square-integrable-martingale for each N and k, the proof now follows similarly to the proof of Lemma 3.7 of [28] and will not be included. Again, the interested reader is referred to [28] for further details. □

We may now state the following result.

Proposition B.0.6. *The sequence $\{\tilde{M}_2^N\}$ is tight.*

Proof. The result follows by the decomposition (184) and Lemmas B.0.4 and B.0.5 above. □

We are now ready to give a proof of Proposition 2.4.2. Before doing so, however, we must first recall Lemma 5.2 from [28]. The proof of this result is similar in our case and therefore will not be included for the sake of brevity.

Let $\beta_i(x,y)$ be bounded real-valued Borel functions such that $E[\beta_i(x,\eta_i)=0]$ and define the processes by $R_m^N=\{R_m^N(t),t\geq 0\}$ and $\langle R_m^N\rangle=\{\langle R_m^N\rangle(t),t\geq 0\}$, $m=1,2,...$, by

$$R_m^N(t)=\sum_{i=1}^{\hat{A}^N(t)\wedge m}\beta_i(\hat{\tau}_i^N,\eta_i)\qquad\text{and}\qquad\langle R_m^N\rangle(t)=\sum_{i=1}^{\hat{A}^N(t)\wedge m}\bar{\beta}_i(\hat{\tau}_i^N),\tag{185}$$

where

$$\bar{\beta}_i(x)\;\;=\;\;E\beta_i^2(x,\eta_i).$$

We also introduce the σ-fields $\hat{\mathcal{F}}_t^N=\sigma\{\hat{\tau}_i^N,\eta_i,1\leq i\leq\lfloor t\rfloor\}\vee\mathcal{N}$ and $\mathcal{F}_t^N=\sigma\{\hat{\tau}_i^N\wedge\hat{\tau}^N_{\hat{A}^N(t)+1},\eta_{i\wedge\hat{A}^N(t)},i\geq 1\}\vee\mathcal{N}$, and the filtrations $\hat{\mathbf{F}}^N=\{\hat{\mathcal{F}}_t^N,t\geq 0\}$ and $\mathbf{F}^N=\{\mathcal{F}_t^N,t\geq 0\}$.

We then have the following.

Lemma B.0.7. *1. The $\hat{\tau}_i^N$, $i=1,2,...$, are $\mathbf{F}^N$-stopping times, and the following inclusions hold: $\mathcal{F}^N_{\hat{\tau}_i^N}\supset\hat{\mathcal{F}}_{i+1}^N$, $\mathcal{G}_i^N\subset\hat{\mathcal{F}}_i^N$, where $\mathcal{G}_i^N=\sigma\{\mathcal{B}\cap\{\hat{\tau}_i^N>t\},t\geq 0,\mathcal{B}\in\mathcal{F}_t^N\}$;*

2. The process $\hat{A}^N$ is $\mathcal{F}^N$-predictable;

3. The processes R_m^N, $m=1,2,...$, are $\mathbf{F}^N$-square-integrable martingales with the processes $\langle R_m^N\rangle$ as predictable quadratic-variation processes.

Proof. See Lemma 5.2 of [28]. □

We may now give a proof of Proposition 2.4.2.

Proof of Proposition 2.4.2. Our proof will be similar to the proof of Lemma 5.3 of [28] but we restate it here for the sake of completeness. Our first step is to show that the finite dimensional distributions of $(\tilde{M}_2^N,\hat{M}_2^N)$ converge to those of $(\tilde{M}_2,\tilde{M}_2)$. We denote finite dimensional convergence by $\overset{f.d.}{\Rightarrow}$.

Let

$$\tilde{U}^N \quad = \quad \frac{U^N}{\sqrt{N}}$$

and note that by Lemma 3.1 of [28], $\tilde{U}^N \Rightarrow \tilde{U}$ in $D([0,\infty), D[0,1])$ as $N \to \infty$, where U is the Kiefer process. Next, let

$$\tilde{M}^N_{2,k}(t) = \sum_{i=1}^{k} \Box\tilde{U}^N((\hat{A}^N(s^k_{i-1}), F(0)), (\hat{A}^N(s^k_i), F(t-s^k_i))), \tag{186}$$

where the increment

$$\Box\tilde{U}^N((a_1,a_2),(b_1,b_2)) \quad = \quad \tilde{U}^N(b_1,b_2) - \tilde{U}^N(a_1,b_2) - \tilde{U}^N(b_1,a_2) + \tilde{U}^N(a_1,a_2),$$

and the points $0 = s^k_0 < s^k_1 < \ldots < s^k_k = t$ are chosen such that

$$\max_{1\le i\le k} |s^k_i - s^k_{i-1}| \to 0 \quad \text{as} \quad k \to \infty.$$

We also define in analogy,

$$M_{2,k}(t) \quad = \quad \sum_{i=1}^{K} (\Box\tilde{U}((e(s^k_{i-1}), F(0)), (e(s^k_i), F(t-s^k_i))) + (\tilde{U}(e(s^k_i), F(0)) - \tilde{U}(e(s^k_{i-1}), F(0)))),$$

where

$$\Box\tilde{U}((a_1,a_2),(b_1,b_2)) \quad = \quad \tilde{U}(b_1,b_2) - \tilde{U}(a_1,b_2) - \tilde{U}(b_1,a_2) + \tilde{U}(a_1,a_2).$$

We will show that

(a) $\tilde{M}^N_{2,k} \stackrel{f.d.}{\Rightarrow} M_{2,k}$,

(b) $\lim_{k\to\infty} \limsup_{N\to\infty} P(|\tilde{M}^N_{2,k}(t) - \tilde{M}^N_2(t)| > \eta) = 0 \quad \text{for} \quad \eta > 0,\ t > 0,$

(c) $\lim_{n\to\infty} P(|\hat{M}^N_2(t) - \tilde{M}^N_2(t)| > \eta) = 0 \quad \text{for} \quad \eta > 0,\ t > 0.$

Since $M_{2,k}(t) \stackrel{P}{\Rightarrow} M_2(t)$ as $k \to \infty$ by definition, this will proved the required.

The proofs of (a) and (b) are identical to the proofs in Lemma 5.3 of [28] but we include them here for the sake of completeness. We proceed as follows.

Thus, by the continuity of the Keifer process $\tilde{U}$, it follows that

$$\check{M}^N_{2,k} \Rightarrow M_{2,k} \quad \text{as} \quad N \to \infty,$$

where

$$\check{M}^N_{2,k}(t) \;=\; \sum_{i=1}^{k} \Box\tilde{U}^N(e(s_i^K), F(0)), (e(s_i^k), F(t-s_i^k))). \tag{187}$$

Next, by Lemma 2.4.1, Lemma 3.1 of [28] and the continuity of U and e, we obtain from (186) and (187) that

$$\lim_{N\to\infty} P\left(\sup_{0\le t\le T} |\tilde{M}^N_{2,k}(t) - \check{M}^N_{2,k}(t)| > \varepsilon\right) = 0, \quad T>0,\ \varepsilon>0. \tag{188}$$

This then implies that $\tilde{M}^N_{2,k} \Rightarrow M_{2,k}$ as $N\to\infty$, which completes the proof of (a).

We will next prove (b), making use of Lemma B.0.7. In the conditions of the lemma we take, fixing t and k for the moment,

$$\beta_i(x.y) \;=\; \sum_{p=1}^{k} 1\{s^k_{p-1} < x \le s^k_p\}(1\{t-s^k_p < x < t-x) - (F(t-x) - F(t-s^k_p))).$$

Then,

$$\begin{aligned}
\bar{\beta}_i(x) &= E[\beta_i(x,\eta_i)^2] \\
&= \sum_{p=1}^{k} 1\{s^k_{p-1} < x \le s^k_p\}(F(t-x) - F(t-s^k_p)) \\
&= \times(1 - F(t-x) - F(t-s^k_p)))
\end{aligned}$$

and (185) yields, by (168), (171) and (186),

$$N^{-1/2}R^N_m(t) \;=\; \tilde{M}^N_2(t) - \tilde{M}^N_{2,k}(t) \ \text{ on } \ \{\hat{A}^N(t)\le m\}. \tag{189}$$

By (189) and (185),

$$\begin{aligned}
N^{-1}\langle R^N_m\rangle(t) &\le N^{-1}\sum_{i=1}^{\hat{A}^N(t)}\sum_{p=1}^{k} 1\{s^k_{p-1} < \hat{\tau}^N_i \le s^k_p\}(F(t-s^k_{p-1}) - F(t-s^k+p)) \\
&= N^{-1}\sum_{p=1}^{k}(F(t-s^k_{p-1}) - F(t-s^k+p))(\hat{A}^N(s^k_p) - \hat{A}^N(s^k_{p-1})) \\
&\le \sup_{1\le p\le k}(N^{-1}\hat{A}^N(s^k_p) - N^{-1}\hat{A}^N(s^k_{p-1})).
\end{aligned}$$

Then, by Lemma B.0.7(3), applying the Lenglart-Rebolledo inequality and (189), for $\eta >$

$0,\ \varepsilon > 0,$

$$\begin{aligned} & P(|\tilde{M}_2^N(t) - \hat{M}_{2,k}^N(t)| > \eta) \\ \leq\ & P(\hat{A}^N(t) > mN) + P(N^{-1/2}|R_m^N(t)| > \eta) \\ \leq\ & P(N^{-1}\hat{A}^N(t) > m) + \frac{\varepsilon}{\eta^2} + P\left(\sup_{1\leq p\leq k}(N^{-1}\hat{A}^N(s_p^k) - N^{-1}\hat{A}^N(s_{p-1}^k)) > \varepsilon\right). \end{aligned}$$

By Lemma 2.4.1, continuity of the indentity function $e(t) = t$ and the fact that $\max_{1\leq p\leq k}(s^k + p - s_{p-1}^k) \to 0$ as $k \to \infty$,

$$\begin{aligned} \lim_{m\to\infty}\limsup_{N\to\infty} P(N^1\hat{A}^N(t) > m) &= 0, \\ \lim_{k\to\infty}\limsup_{N\to\infty} P\left(\sup_{1\leq p\leq k}(N^{-1}\hat{A}^N(s_p^k) - N^{-1}\hat{A}^N(s_{p-1}^k)) > \varepsilon\right) &= 0, \end{aligned}$$

ending the proof of (b).

We next prove part (c). The proof proceeds in a similar manner to the proof of parts (a) and (b). Letting $B^N(t) = \lfloor Nt \rfloor$, we first note that

$$\hat{M}_2^N(t) = N^{-1/2}\int_0^t\int_0^t 1\{s + x \leq t\}dU^N(B^N(s), x).$$

Furthermore, setting $\bar{B}^N = \{N^{-1}B^N(t), t \geq 0\}$, it is clear that

$$\bar{B}^N \Rightarrow e \ \text{ as } \ N \to \infty. \tag{190}$$

Next, letting

$$\check{B}_{2,k}^N(t) = \sum_{i=1}^k \square\tilde{U}^N(\hat{B}^N(s_{i-1}^k), 0), (B^N(s_i^k), F(t - s_i^k)))),$$

it follows by (190), Lemma 3.1 of [28] and the continuity of U and the identity function $e(t) = t$, that

$$\lim_{N\to\infty} P\left(\sup_{0\leq t\leq T}|\check{B}_{2,k}^N(t) - \check{M}_{2,k}^N(t)| > \varepsilon\right) = 0, \quad T > 0,\ \varepsilon > 0. \tag{191}$$

A similar proof to that of part (b) above can also be used to show that

$$\lim_{k\to\infty}\limsup_{N\to\infty} P(|\check{B}_{2,k}^N(t) - \hat{M}_2^N(t)| > \eta) = 0 \ \text{ for } \ \eta > 0,\ t > 0. \tag{192}$$

Part (b), (188), (191) and (192) above now imply part (c).

Parts (a), (b) and (c) imply the finite dimensional convergence $(\tilde{M}_2^N, \hat{M}_2^N) \Rightarrow^{df} (\tilde{M}_2, \tilde{M}_2)$ as $N \to \infty$. It therefore remains to show that the sequence $\{(\tilde{M}_2^N, \hat{M}_2^N)\}$ is tight in order to complete the proof. However, by Proposition B.0.6, the sequence $\{\tilde{M}_2^N\}$ is tight and a similar if not identical proof also shows that $\{\hat{M}_2^N\}$ is tight. Thus, the sequence $\{(\tilde{M}_2^N, \hat{M}_2^N)\}$ is tight, which completes the proof. □

APPENDIX C

CUSTOMER ABANDONMENT PROOFS

In this appendix, we prove Lemmas 3.4.2- 3.5.4 of Chapter III. It is useful for the proof of Lemma 3.4.4 to define the following notation, which matches that in Billingsley [4]. For any set $S \subset [0,T], \delta > 0$, and $x \in D([0,\infty), \Re)$, let

$$w(x,S) \equiv \sup_{u,v \in S} |x(u) - x(v)| \tag{193}$$

$$w_T(x,\delta) \equiv \sup_{0 \le t \le T-\delta} w(x,[t,t+\delta]). \tag{194}$$

Also let

$$w'_T(x,\delta) \equiv \inf \max_{1 \le i \le v} w\left(x,[t_{i-1},t_i)\right),$$

where the infimum extends over all decompositions $[t_{i-1},t_i)$, $1 \le i \le v$, of $[0,T)$ such that $t_i - t_{i-1} > \delta$ for $1 \le i < v$. Similar to (12.7) in [4], since $[0,T)$ can, for each $\delta < \frac{T}{2}$ be split into subintervals $[t_{i-1},t_i)$ satisfying $\delta < t_i - t_{i-1} \le 2\delta$, for $x \in D([0,\infty),\Re)$,

$$w'_T(x,\delta) \le w_T(x,2\delta),\ \delta < \frac{T}{2}. \tag{195}$$

C.1 Proofs of Lemmas 3.4.2 through 3.5.4

Proof of Lemma 3.4.2: Given any $T > 0$, suppose we can show

$$\lim_{n \to \infty} E\left[\sup_{0 \le t \le T} \overline{R}^n(t)\right] = 0. \tag{196}$$

Convergence in L_1 implies convergence in probability, and so

$$\sup_{0 \le t \le T} \overline{R}^n(t) \to 0$$

in probability, as $n \to \infty$. Convergence in probability implies weak convergence, and so

$$\sup_{0 \le t \le T} \overline{R}^n \Rightarrow 0,$$

as $n \to \infty$, which establishes the desired result.

To establish (196), we must show that for any $\delta > 0$ and all large enough n,

$$E\left[\sup_{0\le t\le T} \overline{R}^n(t)\right] < \delta. \tag{197}$$

First observe, using the linearity of the expectation operator and the definitions of $\tilde{V}^n$ in (80) and $\overline{R}^n$ in (117), that

$$E\left[\sup_{0\le t\le T} \overline{R}^n(t)\right] = n^{-1}\sum_{j=1}^{\lfloor nT\rfloor} P\left(\tilde{V}^n(t_j^{n,-}) \ge \sqrt{n}a_j^n\right). \tag{198}$$

Next, we claim there exists a K such that for all n large enough,

$$P\left(\max_{j=1,\ldots,\lfloor nT\rfloor} \tilde{V}^n\left(t_j^{n,-}\right) \ge K\right) < \frac{\delta}{2T}. \tag{199}$$

To see (199), construct a second single server queue on the same probability space as the original queue with abandonments, and with the same arrival and service time sequence as the queue with abandonments, but from which no abandonments occur; i.e., $a_i = \infty$ for all $i \in \{1, 2, \ldots\}$. On a sample path basis, the offered waiting time process in the queue without abandonments always exceeds or is equal to the equivalent process in the queue with abandonments. Weak convergence of a process χ^n in $D([0,\infty), \Re)$ implies tightness of the sequence of random variables $\sup_{0\le t\le T}|\chi^n(t)|$. Therefore, it follows from the weak convergence of the waiting time process for a GI/GI/1 queue established in Theorem 1 in Section 3.2 of Reiman [44] that there exists n_0 such that

$$P\left(\sup_{0\le t\le T}\left|\tilde{V}^n(t)\right| \ge K\right) < \frac{\delta}{2T},\ n \ge n_0 \tag{200}$$

holds. Since for all $j \in \{1, \ldots, \lfloor nT\rfloor\}$, by the definition of t_j^n, the strong law of large numbers, and because assumption (71) implies $\rho^n \to 1$ as $n \to \infty$,

$$t_j^n \le t_{\lfloor nT\rfloor}^n = \frac{\lfloor nT\rfloor}{n\rho^n}\frac{1}{\lfloor nT\rfloor}\sum_{j=1}^{\lfloor nT\rfloor} u_j \to T,$$

as $n \to \infty$, and so (199) holds for large enough n. Finally, recalling the definition of F^n in (74), because $F^n\left(n^{-1/2}K\right) \to 0$ as $n \to \infty$, we can choose n large enough so that

$$F^n\left(\frac{K}{\sqrt{n}}\right) < \frac{\delta}{2T}. \tag{201}$$

Therefore, for any $j \in \{1, \ldots, \lfloor nT \rfloor\}$, from (199) and (201),

$$\begin{aligned} P\left(\tilde{V}^n\left(t_j^{n,-}\right) \geq \sqrt{n}a_j^n\right) &< \frac{\delta}{2T} + P\left(\tilde{V}^n\left(t_j^{n,-}\right) \geq \sqrt{n}a_j^n \cap \max_{j=1,\ldots,\lfloor nT \rfloor} \tilde{V}^n\left(t_j^{n,-}\right) < K\right) \\ &\leq \frac{\delta}{2T} + F^n\left(\frac{K}{\sqrt{n}}\right) < \frac{\delta}{T}, \end{aligned}$$

and so from (198), for large enough n,

$$E\left[\sup_{0 \leq t \leq T} \overline{R}^n(t)\right] < n^{-1} \sum_{j=1}^{\lfloor nT \rfloor} \frac{\delta}{T} \leq \delta,$$

which establishes (197). □

Proof of Lemma 3.4.3: For any given $t, \epsilon, \delta > 0$, we must show

$$P\left(\sup_{0 \leq s \leq t} \left|\tilde{M}_a^n(s)\right| > \epsilon\right) = P\left(\max_{i=1,\ldots,\lfloor nt \rfloor} |M_a^n(i)| > \epsilon\sqrt{n}\right) < \delta, \tag{202}$$

for large enough n. By a generalization of Kolmogorov's inequality (see, for example, Corollary 2.1 in Hall and Heyde [16]), and the orthogonality of martingale differences (see, for example, property (vii) on page 355 of Resnick [45]),

$$\begin{aligned} & P\left(\max_{i=1,\ldots,\lfloor nt \rfloor} |M_a^n(i)| > \epsilon\sqrt{n}\right) \\ \leq & \frac{E\left|M_a^n\left(\lfloor nt \rfloor\right)\right|^2}{\epsilon^2 n} \\ = & \frac{1}{\epsilon^2 n} E\left[\sum_{j=1}^{\lfloor nt \rfloor} \left(\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\} - E\left[\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\} | \mathcal{F}_{j-1}\right]\right)^2\right]. \end{aligned} \tag{203}$$

Since

$$\begin{aligned} \mathbf{1}^2\{V^n\left(t_j^{n,-}\right) \geq a_j^n\} &= \mathbf{1}\{V^n\left(t_j^{n,-}\right) \geq a_j^n\} \\ \mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n\right\} E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n\right\} | \mathcal{F}_{j-1}\right] &\geq 0 \\ E^2\left[\mathbf{1}\{V^n\left(t_j^{n,-}\right) \geq a_j^n\}\right] &\leq E\left[\mathbf{1}^2\{V^n\left(t_j^{n,-}\right) \geq a_j^n\}\right], \end{aligned}$$

it follows that

$$\begin{aligned} & \left(\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\} - E\left[\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\} | \mathcal{F}_{j-1}\right]\right)^2 \\ & \quad \leq \mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\} + E\left[\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\} | \mathcal{F}_{j-1}\right]. \end{aligned} \tag{204}$$

Furthermore,

$$E\left[\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\} + E\left[\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\}|\mathcal{F}_{j-1}\right]\right] = 2E\left[\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\}\right], \tag{205}$$

and so from (203), (204), (205), and the definition of $\overline{R}^n$ in (117),

$$P\left(\max_{i=1,\ldots,\lfloor nt\rfloor} |M_a^n(i)| > \epsilon\sqrt{n}\right) \leq \frac{2}{\epsilon^2 n}\sum_{j=1}^{\lfloor nt\rfloor} E\left[\mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n\right\}\right] = \frac{2}{\epsilon^2}E\left[\overline{R}^n(t)\right] \to 0,$$

as $n \to \infty$, by the convergence established in (196). □

Proof of Lemma 3.4.4: We first argue that the families $\{\tilde{X}^n\}$ and $\{\tilde{\epsilon}^n\}$ are tight in $D([0,\infty),\Re)$, and then use those tightness results to establish the tightness of $\{\tilde{V}^n\}$.

Tightness of $\{\tilde{\mathbf{X}}^{\mathbf{n}}\}$:

We first argue

$$\tilde{S}_a^n \circ \overline{A}^n \Rightarrow 0, \tag{206}$$

as $n \to \infty$. Recall the distributional equivalence in (119)

$$\tilde{S}^n \circ \overline{R}^n \circ \overline{A}^n \stackrel{D}{=} \tilde{S}_a^n \circ \overline{A}^n.$$

Hence, it is sufficient to show

$$\tilde{S}^n \circ \overline{R}^n \circ \overline{A}^n \Rightarrow 0,$$

as $n \to \infty$. From the weak convergence in (87) and Lemma 3.4.2,

$$\left(\tilde{S}^n, \overline{R}^n\right) \Rightarrow \left(\mathrm{var}(v_1)W_{S,2}, 0\right),$$

as $n \to \infty$. The joint convergence holds because $\overline{R}^n$ weakly converges to a constant. Since $\overline{R}^n$ and $\overline{A}^n$ are non-decreasing in t for each n, the almost sure convergence of $\overline{A}^n$ in (86), and the random time change theorem imply

$$\tilde{S}^n \circ \overline{R}^n \circ \overline{A}^n \Rightarrow 0,$$

as $n \to \infty$.

From the definition of $\tilde{X}^n$ in (111) and the evolution equation for X^n in (91),

$$\tilde{X}^n = \tilde{A}^n + \tilde{S}^n(\bar{A}^n) + \sqrt{n}t(\rho^n - 1) - \tilde{S}_a^n(\bar{A}^n(t)) - \tilde{M}_a^n(\bar{A}^n).$$

The weak convergences in (87), the heavy traffic assumption (71), the almost sure convergence of $\overline{A}^n$ in (86), Lemma 3.4.3, the weak convergence in (206), and the random time change theorem imply that

$$\tilde{X}^n \Rightarrow W,$$

as $n \to \infty$, where W is a Brownian motion with drift θ and variance $\sigma^2 = \text{var}(u_1) + \text{var}(u_2)$. Tightness follows because weakly convergence subsequences are relatively compact.

Tightness of $\{\tilde{\epsilon}^n\}$:

Let $T > 0$. We verify the conditions (16.17) and (16.18) of Theorem 16.8 in Billingsley [4] to prove the tightness of $\{\tilde{\epsilon}^n\}$. In particular, we must show the following.

- **(B16.17)** For every $\eta > 0$, there exists an a and an n_0 such that

$$P\left(\sup_{0 \le t \le T} |\tilde{\epsilon}^n(t)| \ge a\right) < \eta, \ n \ge n_0;$$

- **(B16.18)** For every γ and η, there exists a δ and an n_0 such that

$$P\left(w_T'(\tilde{\epsilon}^n, \delta) \ge \gamma\right) < \eta, \ n \ge n_0.$$

To see conditions (B16.17) and (B16.18) can be satisfied, first recall from (124) that

$$\tilde{\epsilon}^n(t) = \int_0^t \left(\int_0^{\tilde{V}^n(s^-)} h(w)dw\right) ds - \int_0^t \sqrt{n}\left(1 - \exp\left(-\frac{1}{\sqrt{n}}\int_0^{\tilde{V}^n(s^-)} h(w)dw\right)\right) d\overline{A}^n(s). \tag{207}$$

Choose K, a, and n_0 large enough, and δ small enough so that

$$P\left(\sup_{0 \le t \le T} \left|\tilde{V}^n(t)\right| \ge K\right) < \frac{\eta}{2} \tag{208}$$

$$P\left(K\|h\|_K T + (K\|h\|_K + 1)\,\overline{A}^n(T) \ge a\right) < \frac{\eta}{2} \tag{209}$$

$$P\left(K\|h\|_K 2\delta + (1 + K\|h\|_K)\left(\overline{A}^n(t + 2\delta) - \overline{A}^n(t)\right) \ge \gamma\right) < \frac{\eta}{2}, \tag{210}$$

for all $n \geq n_0$, where $\|h\|_K < \infty$ because h is continuous. Such a K, a, δ, and n_0 exist from the observation (200) in the proof of Lemma 3.4.2, and the almost sure convergence of $\overline{A}^n$ to the identity function in (86). Then, from (207), (208), and (209), also noting that

$$\sqrt{n}\left(1 - \exp\left(-x/\sqrt{n}\right)\right) \to x, \tag{211}$$

uniformly on compact sets of $[0, \infty)$,

$$\begin{aligned} P\left(\sup_{0\leq t\leq T} |\tilde{\epsilon}^n(t)| \geq a\right) &\leq \frac{\eta}{2} + P\left(\sup_{0\leq t\leq T} |\tilde{\epsilon}^n(t)| \geq a \cap \sup_{0\leq t\leq T} \left|\tilde{V}^n(t)\right| < K\right) \\ &\leq \frac{\eta}{2} + P\left(TK\|h\|_K + (K\|h\|_K + 1)\,\overline{A}^n(T) \geq a\right) < \eta, \end{aligned}$$

and so (B16.17) holds. Next, from the definitions of w and w_T in (193) and (194), the inequality (195), and the non-negativity of h,

$$\begin{aligned} w'_T\left(\tilde{\epsilon}^n, \delta\right) &\leq w_T\left(\tilde{\epsilon}^n, 2\delta\right) \\ &\leq \sup_{0\leq t\leq T-2\delta} \int_t^{t+2\delta} \left(\int_0^{\tilde{V}^n(s^-)} h(w)dw\right) ds \\ &\quad + \int_t^{t+2\delta} \sqrt{n}\left(1 - \exp\left(-\frac{1}{\sqrt{n}} \int_0^{\tilde{V}^n(s^-)} h(w)dw\right)\right) d\overline{A}^n(s), \end{aligned}$$

and so, also using (208), (210), and (211),

$$\begin{aligned} P\left(w'_T\left(\tilde{\epsilon}^n, \delta\right) \geq \gamma\right) &\leq \frac{\eta}{2} + P\left(w'_T\left(\tilde{\epsilon}^n, \delta\right) \geq \gamma \cap \sup_{0\leq t\leq T} \left|\tilde{V}^n(t)\right| \leq K\right) \\ &\leq \frac{\eta}{2} + P\left(2\delta K\|h\|_K + (1 + K\|h\|_K)\left(\overline{A}^n(t + 2\delta) - \overline{A}^n(t)\right) \geq \gamma\right) \\ &< \eta, \end{aligned}$$

which implies (B16.18) holds. We conclude $\{\tilde{\epsilon}^n\}$ is tight.

Tightness of $\{\tilde{\mathbf{V}}^{\mathbf{n}}\}$:

We show the sequence $\{\tilde{V}^n\}$ satisfies the definition of relative compactness, and so is tight in $D([0, \infty), \Re)$. Consider any subsequence $\{\tilde{V}^{n_i}\}$. Because the families $\{\tilde{X}^n\}$ and $\{\tilde{\epsilon}^n\}$ are both tight, there exists a further subsequence $\{\tilde{X}^{n_i(m)} + \tilde{\epsilon}^{n_i(m)}\}$ such that

$$\tilde{X}^{n_i(m)} + \tilde{\epsilon}^{n_i(m)} \Rightarrow \chi,$$

as $n_i(m) \to \infty$, for some limit process χ. From the representation for $\tilde{V}^n$ in (114), the continuous mapping theorem, and the continuity of the mapping ϕ^h established in part (iii) of Proposition 3.3.3, on this further subsequence,

$$\tilde{V}^{n_i(m)} = \phi^h\left(\tilde{X}^{n_i(m)} + \tilde{\epsilon}^{n_i(m)}\right) \Rightarrow \phi^h(\chi),$$

as $n_i(m) \to \infty$. We conclude $\{\tilde{V}^n\}$ is relatively compact. □

Proof of Lemma 3.4.5: The offered waiting time process can only increase at arrival time points and so

$$V^n(t) \leq \max_{i=1,\ldots,A^n(t)} V^n(t_i). \tag{212}$$

Since in the n^{th} system service times are scaled by n^{-1} and the service time of the ith arrival is only included in the offered waiting time process if $V^n(t_i) \leq C^n$, recalling the definition for C^n in (75), we find

$$\max_{i=1,\ldots,A^n(t)} V^n(t_i) \leq \frac{C}{\sqrt{n}} + \max_{i=1,\ldots,A^n(t)} \frac{v_i}{n}. \tag{213}$$

From (213), the fact that the v_i's are non-negative random variables, the definition of δ^n in (129), and (212)

$$\sup_{0\leq s\leq t} \sqrt{n}\delta^n(s) = \max_{i=1,\ldots,A^n(t)} \left[\tilde{V}^n(t_i) - C\right]^+ \leq \max_{i=1,\ldots,A^n(t)} \frac{v_i}{\sqrt{n}}. \tag{214}$$

Since $\sqrt{n}\delta^n$ is a non-negative process, $n^{-1}A^n \to e$ as $n \to \infty$, almost surely, uniformly on compact sets, and Lemma 3.3 in Iglehart and Whitt [20] establishes $\max_{i=1,\ldots,nt} n^{-1/2}v_i \Rightarrow 0$ as $n \to \infty$, the random time change theorem and (214) imply

$$\sqrt{n}\delta^n \Rightarrow 0,$$

as $n \to \infty$. □

It is useful for the proof of Lemma 3.4.6 to observe that

$$(\mathcal{V}^n, I^n, \mathcal{U}^n) = (\phi_{C^n}, \psi_{1,C^n}, \psi_{2,C^n})\,(\chi^n - \delta^n)\,, \tag{215}$$

where

$$\begin{aligned} \chi^n(t) &= \frac{1}{n}A^n(t) - \rho^n t + S^n\left(A^n(t)\right) + t\left(\rho^n - 1\right) \\ &\quad -S_a^n(A^n(t)) - \frac{1}{n}\sum_{j=1}^{A^n(t)} \mathbf{1}\left\{V^n(t_j^{n,-}) \geq a_j^n \cap \tilde{V}^n\left(t_j^{n,-}\right) < C\right\}, \end{aligned} \tag{216}$$

and

$$\mathcal{U}^n(t) = \frac{1}{n} \sum_{j=1}^{A^n(t)} \mathbf{1}\left\{\tilde{V}^n\left(t_j^{n,-}\right) \geq C\right\}. \tag{217}$$

To see (215) is valid, we verify the conditions (C1) and (C2) of Definition 3.3.4 when h is the zero function.

(C1) From (127) and the fact that the process V^n is non-negative, $0 \leq \mathcal{V}^n \leq C^n$. We now show

$$\mathcal{V}^n = \chi^n - \delta^n + I^n - \mathcal{U}^n.$$

From (93)-(95),

$$\begin{aligned} V^n(t) &= X^n(t) + \epsilon_B^n(t) - \int_0^t \left(\int_0^{V^n(s^-) \wedge C^n} h^n(u) du \right) ds + I^n(t) - U^n(t) \\ &= X^n(t) - \frac{1}{n} \int_0^t F^n\left(V^n(s^-)\right) dA^n(s) + I^n(t). \end{aligned}$$

The definitions of X^n in (91) and M_a in (64) then imply

$$\begin{aligned} V^n(t) &= \frac{1}{n} A^n(t) - \rho^n t + S^n\left(A^n(t)\right) - S_a^n\left(A^n(t)\right) \\ &\quad + t\left(\rho^n - 1\right) - \frac{1}{n} \sum_{j=1}^{A^n(t)} \mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n\right\} + I^n(t). \end{aligned}$$

Since the abandonment times $\{a_j^n, j = 0, 1, 2, \ldots\}$ are bounded above by $\sqrt{n}C$,

$$\mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n\right\} = \mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n \cap \tilde{V}^n\left(t_j^{n,-}\right) < C\right\} + \mathbf{1}\left\{\tilde{V}^n\left(t_j^{n,-}\right) \geq C\right\},$$

and so

$$V^n(t) = \chi^n(t) + I^n(t) - \mathcal{U}^n(t).$$

From (128), $\mathcal{V}^n = V^n - \delta^n$, and so

$$\mathcal{V}^n(t) = \chi^n(t) - \delta^n(t) + I^n(t) - \mathcal{U}^n(t).$$

(C2) The processes I^n and $\mathcal{U}^n$ are non-decreasing functions having $I^n(0) = \mathcal{U}^n(0) = 0$, and

$$\int_0^\infty \mathcal{V}^n(t) dI^n(t) = \int_0^\infty \left(V^n(t) \wedge C^n\right) \mathbf{1}\left\{V^n(t) = 0\right\} = 0,$$

and

$$\begin{aligned}
&\int_0^\infty [C^n - \mathcal{V}^n(t)]^+ \, d\mathcal{U}^n(t) \\
= \;& \frac{1}{n} \int_0^\infty [C^n - (V^n(t) \wedge C^n)]^+ \, d \left(\sum_{j=1}^{A^n(t)} \mathbf{1} \left\{ V^n \left(t_j^{n,-} \right) \geq C^n \right\} \right) = 0.
\end{aligned}$$

Proof of Lemma 3.4.6: Because $C^n = n^{-1/2}C$ upper bounds the abandonment times $\{a_j^n, j = 0, 1, 2, \ldots\}$,

$$\overline{R}^n(t) = n^{-1} \sum_{j=1}^{\lfloor nt \rfloor} \mathbf{1}\{V^n \left(t_j^{n,-} \right) \geq a_j^n \cap \tilde{V}^n \left(t_j^{n,-} \right) < C\} + n^{-1} \sum_{j=1}^{\lfloor nt \rfloor} \mathbf{1}\{\tilde{V}^n \left(t_j^{n,-} \right) \geq C\} \quad (218)$$

Using the expression for F^n in (76),

$$\begin{aligned}
& E \left[\sup_{0 \leq t \leq T} n^{-1} \sum_{j=1}^{\lfloor nt \rfloor} \mathbf{1}\{V^n \left(t_j^{n,-} \right) \geq a_j^n \cap \tilde{V}^n \left(t_j^{n,-} \right) < C\} \right] \\
= \;& n^{-1} \sum_{j=1}^{\lfloor nT \rfloor} P \left(V^n \left(t_j^{n,-} \right) \geq a_j^n \cap \tilde{V}^n \left(t_j^{n,-} \right) < C \right) \\
\leq \;& n^{-1} \lfloor nT \rfloor \left(1 - \exp \left(\frac{-1}{\sqrt{n}} \int_0^C h(w) dw \right) \right) \\
\to \;& 0,
\end{aligned}$$

as $n \to \infty$. Convergence in L_1 implies convergence in probability, and so

$$\sup_{0 \leq t \leq T} n^{-1} \sum_{j=1}^{\lfloor nt \rfloor} \mathbf{1} \left\{ V^n \left(t_j^{n,-} \right) \geq a_j^n \cap \tilde{V}^n \left(t_j^{n,-} \right) < C \right\} \to 0, \quad (219)$$

in probability, as $n \to \infty$.

For the second term in (218), we first observe from (79), (81), and (82), and (216) that

$$\begin{aligned}
\sqrt{n}\chi^n(t) \;=\; & \tilde{A}^n(t) + \tilde{S}^n \left(\overline{A}^n(t) \right) + \sqrt{n}t \left(\rho^n - 1 \right) \\
& - \tilde{S}_a^n \left(\overline{A}^n(t) \right) - \frac{1}{\sqrt{n}} \sum_{j=1}^{A^n(t)} \mathbf{1} \left\{ V^n \left(t_j^{n,-} \right) \geq a_j^n \cap \tilde{V}^n \left(t_j^{n,-} \right) < C \right\}.
\end{aligned}$$

Let W be a Brownian motion with drift θ and variance $\sigma^2 = \mathrm{var}(u_1) + \mathrm{var}(v_1)$. The almost sure convergence in (86), the weak convergence in (87), the random time change theorem, and the heavy traffic assumption in (71) establish the weak convergence of the first three terms in the above expression for $\sqrt{n}\chi^n$ to W. The arguments to show the weak convergence

of the fourth term to 0 in (206) in the proof of Lemma 3.4.4 remain valid. To see the fifth term weakly converges to zero, suppose we can show

$$n^{-1/2} \sum_{j=1}^{A^n(\cdot)} \mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\} \Rightarrow \int_0^{\cdot} h(u)du, \tag{220}$$

as $n \to \infty$, which implies the process on the left-hand side is tight in $D([0,\infty), \Re)$. Then, because for each $s \leq t$,

$$\sum_{j=A^n(s)}^{A^n(t)} \mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n \cap \tilde{V}^n\left(t_j^{n,-}\right) < C\right\} \leq \sum_{j=A^n(s)}^{A^n(t)} \mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\},$$

it follows from Theorem 16.8 in Billingsley (which provides sufficient conditions for tightness in $D([0,\infty), \Re)$ that

$$\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{A^n(t)} \mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n \cap \tilde{V}^n\left(t_j^{n,-}\right) < C\right\} \right\}$$

is tight in $D([0,\infty), \Re)$. Consider any subsequence n_k on which

$$\frac{1}{\sqrt{n_k}} \sum_{j=0}^{A^{n_k}(\cdot)} \mathbf{1}\left\{V^n\left(t_j^{n_k,-}\right) \geq a_j^{n_k} \cap \tilde{V}^{n_k}\left(t_j^{n_k,-}\right) < C\right\} \Rightarrow \Psi,$$

as $n_k \to \infty$. On this subsequence, the two-sided conventional regulator mapping representation in (215), the scaling property in part (ii) of Proposition 3.3.6, Lemma 3.4.5, the continuous mapping theorem, and the continuity of $\psi_{2,C}$ established in Theorem 14.8.1 in Whitt [54] imply

$$\sqrt{n_k}\mathcal{U}^{n_k} = \psi_{2,C}\left(\sqrt{n_k}\chi^{n_k} - \delta^{n_k}\right) \Rightarrow \psi_{2,C}\left(W + \Psi\right),$$

as $n_k \to \infty$, and so

$$\mathcal{U}^{n_k} \Rightarrow 0,$$

as $n_k \to \infty$. Since the subsequence n_k was arbitrary,

$$\mathcal{U}^{n} \Rightarrow 0,$$

as $n \to \infty$. Finally, from (218), (219), the random time change theorem, and the fact that convergence in probability implies weak convergence,

$$\overline{R}^n \Rightarrow 0,$$

as $n \to \infty$.

Weak Convergence of $n^{-1/2} \sum_{j=1}^{A^n(\cdot)} \mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\}$:

Consider the centered sum

$$\begin{aligned} \triangle^n(t) &\equiv \frac{1}{n}\sum_{j=1}^{\lfloor nt \rfloor} \left(\sqrt{n}\mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\} - E\left[\sqrt{n}\mathbf{1}\{a_j^n < C/\sqrt{n}\}\right]\right) \\ &= \left(\frac{1}{n}\sum_{j=1}^{\lfloor nt \rfloor} \sqrt{n}\mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\}\right) - \frac{\lfloor nt \rfloor}{n}\sqrt{n}P(a_1^n < C/\sqrt{n}). \end{aligned}$$

From the representation for F^n in (76) and L'Hopital's rule, we have that

$$\sqrt{n}P(a_1^n < C/\sqrt{n}) = \sqrt{n}\left(1 - \exp\left(-\frac{1}{\sqrt{n}}\int_0^C h(u)du\right)\right) \to \int_0^C h(u)du, \tag{221}$$

as $n \to \infty$, and so

$$\lim_{n\to\infty} \frac{\lfloor nt \rfloor}{n}\sqrt{n}P\left(a_1^n \le C/\sqrt{n}\right) = t\int_0^C h(u)du.$$

Therefore, if we can show that $\triangle^n(t) \to 0$ in probability as $n \to \infty$, then it must be true that for each $t \ge 0$,

$$\frac{1}{n}\sum_{j=1}^{\lfloor nt \rfloor} \sqrt{n}\mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\} \to t\int_0^C h(u)du, \tag{222}$$

in probability, as $n \to \infty$. For any $t > 0$, in order to show that $\triangle^n(t) \to 0$ in probability as $n \to \infty$, it is sufficient to show that $E\left[\triangle^n(1)^2\right] \to 0$. (This is because convergence in L^2 implies convergence in probability.) By independence,

$$\begin{aligned} E\left[\triangle^n(t)^2\right] &= \frac{1}{n}\sum_{j=1}^{\lfloor nt \rfloor} E\left[\left(\mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\} - E\left[\mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\}\right]\right)^2\right] \\ &\le \frac{2}{n}\sum_{j=1}^{\lfloor nt \rfloor} E\left[\mathbf{1}\left\{a_j^n < C/\sqrt{n}\right\}\right] \\ &= 2\frac{\lfloor nt \rfloor}{n}P\left(a_j^n < C/\sqrt{n}\right) \\ &\to 0, \end{aligned}$$

as $n \to \infty$ by (221), and so (222) is valid.

The limit point on the right-hand side of (222) is deterministic, and convergence in probability implies weak convergence. Repeated application of Theorem 3.9 in Billingsley [4] then implies that the finite dimensional distributions weakly converge. If we can argue the process on the left-hand side of (222) is tight, then we can conclude the process level convergence

$$\left\{ n^{-1/2} \sum_{j=1}^{\lfloor n \cdot \rfloor} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\} \right\} \Rightarrow \int_0^{\cdot} h(u) du$$

as $n \to \infty$ is valid. The random time change theorem and (86) then imply that the weak convergence in (220) is valid, completing the proof.

We verify conditions (16.17) and (16.18) of Theorem 16.8 in Billingsley [4] to show that the process

$$\left\{ n^{-1/2} \sum_{j=1}^{A^n(\cdot)} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\} \right\}$$

is tight in $D([0,\infty), \Re)$.

(B16.17) For each $a > T \int_0^C h(u)du$, from (222)

$$P\left(\sup_{0 \le t \le T} \left| \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} \sqrt{n} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\} \right| > a \right) = P\left(\frac{1}{n} \sum_{j=1}^{\lfloor nT \rfloor} \sqrt{n} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\} > a \right) \to 0,$$

as $n \to \infty$.

(B16.18) First note that

$$\begin{aligned} & w_T'\left(\frac{1}{n} \sum_{j=1}^{\lfloor n \cdot \rfloor} \sqrt{n} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\}, \delta \right) \\ \le & \max_{i \in \{0,\ldots,\lfloor T/\delta \rfloor + 1\}} \left(\frac{1}{n} \sum_{j=0}^{\lfloor n\delta(i+1) \rfloor} \sqrt{n} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\} - \frac{1}{n} \sum_{j=0}^{\lfloor n\delta i \rfloor} \sqrt{n} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\} \right) \\ \Rightarrow & \delta \int_0^C h(u) du \text{ as } n \to \infty. \end{aligned}$$

The above weak convergence follows by the convergence of the finite dimensional distributions and the continuous mapping theorem. Thus, since convergence in distribution to a constant implies convergence in probability as well, we have that for each

$\varepsilon > 0$,

$$\lim_{\delta \to 0} \lim_{n \to \infty} P\left(w_T^{'} \left(\frac{1}{n} \sum_{j=1}^{\lfloor n \cdot \rfloor} \sqrt{n} \mathbf{1}\left\{ a_j^n < C/\sqrt{n} \right\}, \delta \right) > \varepsilon \right) = 0,$$

as $n \to \infty$.

□

Proof of Lemma 3.4.7: Since $\overline{R}^n \Rightarrow 0$ as $n \to \infty$ under Assumption 2 by Lemma 3.4.6, the arguments to prove Lemma 3.4.3 remain valid. □

Proof of Lemma 3.4.8: By the representation of $\tilde{\mathcal{V}}^n$ in (132) and the continuous mapping theorem, it is sufficient to show the sequences $\left\{ \tilde{X}^n \right\}, \left\{ \tilde{\delta}^n \right\}$, and $\{\tilde{\epsilon}_B^n\}$ are tight in $D\left([0,\infty), \Re\right)$. The sequence $\{\tilde{X}^n\}$ is tight in $D([0,\infty), \Re)$ by the same arguments as in the proof of Lemma 3.4.4. Lemma 3.4.5 establishes the sequence $\{\tilde{\delta}^n\}$ is tight in $D([0,\infty), \Re)$. The following argument shows the sequence $\{\tilde{\epsilon}_B^n\}$ is tight in $D([0,\infty), \Re)$. The evolution equation for $\tilde{\epsilon}_B^n$ in (139) is exactly that for $\tilde{\epsilon}^n$ in (124), with $\tilde{\mathcal{V}}^n$ replacing $\tilde{V}^n$. Therefore, almost the same arguments as in Lemma 3.4.4 can be used to show $\tilde{\epsilon}_B^n$ satisfies conditions (16.17) and (16.18) in Theorem 16.8 in Billingsley [4] (given in (B16.17) and (B16.18) in the proof of Lemma 3.4.4 in terms of $\tilde{\epsilon}^n$), which establishes $\{\tilde{\epsilon}_B^n\}$ is tight in $D([0,\infty), \Re)$. The difference is a simplification: the choice of large K in (208) is not necessary because from the definitions of C^n in (75), $\mathcal{V}^n$ in (127), and $\tilde{\mathcal{V}}^n$ in (132),

$$\tilde{\mathcal{V}}^n(t) \leq C \text{ for all } t \geq 0.$$

□

Proof of Lemma 3.5.1: We first prove Lemma 3.5.1 under Assumption 1 and then under Assumption 2.

Proof under Assumption 1: First observe from the definitions of M_a and $\tilde{M}_a^n$ in (64) and (84) that

$$\begin{aligned} & n^{-1/2} \sum_{i=A^n \circ a^n(t)}^{A^n(t)} \mathbf{1}\left\{ V^n\left(t_i^{n,-} \right) \geq a_i^n \right\} \\ = \; & \tilde{M}_a^n\left(\overline{A}^n(t) \right) - \tilde{M}_a^n\left(\overline{A}^n(a^n(t)) \right) + n^{-1/2} \sum_{i=A^n(a^n(t))}^{A^n(t)} F^n\left(V^n(t_i^{n,-}) \right). \end{aligned}$$

Lemma 3.4.3, the convergence of $\overline{A}^n$ to the identity process in (86), the fact that $a^n(t) \leq t$ for all $t \geq 0$, and the random time change theorem show

$$\tilde{M}_a^n \circ \overline{A}^n - \tilde{M}_a^n \circ \overline{A}^n \circ a^n \Rightarrow 0, \tag{223}$$

as $n \to \infty$. Therefore, to show the stated result under Assumption 1, it remains to show that

$$n^{-1/2} \sum_{i=A^n(a^n(t))}^{A^n(t)} F^n\left(V^n(t_i^{n,-})\right) \Rightarrow 0, \tag{224}$$

as $n \to \infty$.

To show (224), it is sufficient to show that for any $\gamma, \delta, T > 0$,

$$P\left(\sup_{0\leq t\leq T} n^{-1/2} \sum_{i=A^n(a^n(t))}^{A^n(t)} F^n\left(V^n(t_i^{n,-})\right) > \gamma\right) < \delta. \tag{225}$$

For any $\delta > 0$, from (199) and the convergence of $\overline{A}^n$ in (86), we can choose K large enough so that

$$P\left(\max_{j=1,\ldots,\lfloor n\overline{A}^n(t)\rfloor} \tilde{V}^n\left(t_j^{n,-}\right) \geq K\right) < \frac{\delta}{2},$$

and so

$$\begin{aligned}
&P\left(\sup_{0\leq t\leq T} n^{-1/2} \sum_{i=A^n(a^n(t))}^{A^n(t)} F^n\left(V^n(t_i^{n,-})\right) > \gamma\right) \qquad (226)\\
&< \frac{\delta}{2} + P\left(\sup_{0\leq t\leq T} n^{-1/2} \sum_{i=A^n(a^n(t))}^{A^n(t)} F^n\left(V^n(t_i^{n,-})\right) > \gamma \cap \max_{j=1,\ldots,\lfloor n\overline{A}^n(t)\rfloor} \tilde{V}^n(t_j^{n,-}) < K\right)\\
&\leq \frac{\delta}{2} + P\left(\sup_{0\leq t\leq T} \left(A^n(t) - A^n(a^n(t))\right) n^{-1/2} F^n\left(\frac{K}{\sqrt{n}}\right) > \gamma\right)\\
&\leq \frac{\delta}{2} + P\left(\sup_{0\leq t\leq T} \left(\tilde{A}^n(t) - \tilde{A}^n(a^n(t))\right) F^n\left(\frac{K}{\sqrt{n}}\right) + \rho^n(t - a^n(t))\sqrt{n}F^n\left(\frac{K}{\sqrt{n}}\right) > \gamma\right).
\end{aligned}$$

From the definition of F^n in (74) and L'Hopital's rule,

$$\sqrt{n}F^n\left(\frac{K}{\sqrt{n}}\right) = \sqrt{n}\left(1 - \exp\left(-\frac{1}{\sqrt{n}}\int_0^K h(w)dw\right)\right) \to \int_0^K h(w)dw,$$

as $n \to \infty$. The function h is continuous, and so $\sup_{0\leq w\leq K} |h(w)| < \infty$. Furthermore, $\tilde{A}^n$ converges to a continuous limit process. Therefore, the weak convergence of a^n to the identity process e in (141) implies

$$\left(\tilde{A}^n - \tilde{A}^n \circ a^n\right) F^n\left(\frac{K}{\sqrt{n}}\right) + \rho^n\left(e - a^n\right)\sqrt{n}F^n\left(\frac{K}{\sqrt{n}}\right) \Rightarrow 0,$$

as $n \to \infty$. Since weak convergence to a constant is equivalent to convergence in probability, we can choose n large enough so that

$$P\left(\sup_{0\le t\le T}\left(\tilde{A}^n(t)-\tilde{A}^n(a^n(t))\right)F^n\left(\frac{K}{\sqrt{n}}\right)+\rho^n(t-a^n(t))\sqrt{n}F^n\left(\frac{K}{\sqrt{n}}\right)>\gamma\right)<\frac{\delta}{2},$$

which, from (226) implies (225) is valid, and completes the proof under Assumption 1.

Proof under Assumption 2: Define

$$\begin{aligned}
U_A^n(i) &\equiv \frac{1}{n}\sum_{j=1}^{i}\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge C^n\right\}\\
\tilde{U}_A^n(t) &\equiv \sqrt{n}U_A^n(\lfloor nt\rfloor)\\
M^n(i) &\equiv \sum_{j=1}^{i}(\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge a_j^n\cap V^n\left(t_j^{n,-}\right)<C^n\right\}\\
&\quad -E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge a_j^n\cap V^n\left(t_j^{n,-}\right)<C^n\right\}|\mathcal{F}_{j-1}\right])\\
\tilde{M}^n(t) &\equiv \frac{1}{\sqrt{n}}M^n\left(\lfloor nt\rfloor\right),
\end{aligned}$$

and observe that

$$\begin{aligned}
&n^{-1/2}\sum_{i=A^n\circ a^n(t)}^{A^n(t)}\mathbf{1}\left\{V^n\left(t_i^{n,-}\right)\ge a_i^n\right\} \qquad (227)\\
&= n^{-1/2}\sum_{i=A^n\circ a^n(t)}^{A^n(t)}\mathbf{1}\left\{V^n\left(t_i^{n,-}\right)\ge a_i^n\cap V^n(t_i^{n,-})\ge C^n\right\}\\
&\quad +n^{-1/2}\sum_{i=A^n\circ a^n(t)}^{A^n(t)}\mathbf{1}\left\{V^n\left(t_i^{n,-}\right)\ge a_i^n\cap V^n(t_i^{n,-})<C^n\right\}\\
&= n^{-1/2}\sum_{i=A^n\circ a^n(t)}^{A^n(t)}\mathbf{1}\left\{V^n(t_i^{n,-})\ge C^n\right\}+n^{-1/2}\sum_{i=A^n\circ a^n(t)}^{A^n(t)}\mathbf{1}\left\{V^n\left(t_i^{n,-}\right)\ge a_i^n\cap V^n(t_i^{n,-})<C^n\right\}\\
&= \tilde{M}^n\circ\overline{A}^n(t)-\tilde{M}^n\circ\overline{A}^n\circ a^n(t)+\tilde{U}_A^n\circ\overline{A}^n(t)-\tilde{U}_A^n\circ\overline{A}^n\circ a^n(t)\\
&\quad +n^{-1/2}\sum_{i=A^n\circ a^n(t)}^{A^n(t)}E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge a_j^n\cap V^n\left(t_j^{n,-}\right)<C^n\right\}|\mathcal{F}_{j-1}\right].
\end{aligned}$$

We show each of the three terms on the right-hand side of (227) weakly converges to 0 as $n\to\infty$.

Arguments identical to those in the proof of Lemma 3.4.3 show that for any $t>0$

$$P\left(\max_{i=1,\ldots,\lfloor nt\rfloor}|M^n(i)|>\epsilon\sqrt{n}\right)\le\frac{2}{\epsilon^2 n}\sum_{j=1}^{\lfloor nt\rfloor}E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge a_j^n\cap V^n\left(t_j^{n,-}\right)<C^n\right\}|\mathcal{F}_{j-1}\right]. \qquad (228)$$

Since from the expression for F^n in (76), recalling that $C^n = n^{-1/2}C$ from (75),

$$E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n \cap V^n\left(t_j^{n,-}\right) < C^n\right\}|\mathcal{F}_{j-1}\right] \leq 1 - \exp\left(-\frac{1}{\sqrt{n}}\int_0^C h(w)dw\right) \to 0,$$

as $n \to \infty$, (228) implies

$$P\left(\max_{i=1,\ldots,\lfloor nt\rfloor}|M^n(i)| > \epsilon\sqrt{n}\right) \to 0,$$

as $n \to \infty$, and so,

$$\tilde{M}^n \Rightarrow 0,$$

as $n \to \infty$. The almost sure convergence of $\overline{A}^n$ in (86) and the weak convergence of a^n in (141) then imply

$$\tilde{M}^n \circ \overline{A}^n - \tilde{M}^n \circ \overline{A}^n \circ a^n \Rightarrow 0, \tag{229}$$

as $n \to \infty$. Next, because $\tilde{U}_A^n\left(\overline{A}^n(t)\right) = (b^n)^{-1}\tilde{U}^n(t)$ (recalling the definitions of U^n and $\tilde{U}^n$ in (93) and (134)), $b^n \to 1$ as $n \to \infty$, and $\overline{A}^n \to e$ almost surely, uniformly on compact sets, as $n \to \infty$, part (ii) of Theorem 3.4.1 and the random time change theorem establish

$$\tilde{U}_A^n \Rightarrow \psi_{2,C}^h(W),$$

as $n \to \infty$, where W is a Brownian motion as defined in Theorem 3.4.1. Thus, since $\psi_{2,C}^h(W)$ is almost surely a continuos process, this then implies by (141) and the fact that $\overline{A}^n \to e$ as $n \to \infty$, that

$$\tilde{U}_A^n \circ \overline{A}^n - \tilde{U}_A^n \circ \overline{A}^n \circ a^n \Rightarrow 0, \tag{230}$$

as $n \to \infty$. Finally, for the last term on the right-hand side of (227), again using (76),

$$\begin{aligned} & n^{-1/2}\sum_{i=A^n\circ a^n(t)}^{A^n(t)} E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right) \geq a_j^n \cap V^n\left(t_j^{n,-}\right) < C^n\right\}|\mathcal{F}_{j-1}\right] \qquad (231) \\ \leq\ & \left(\overline{A}^n(t) - \overline{A}^n \circ a^n(t)\right)\sqrt{n}\left(1 - \exp\left(\frac{-1}{\sqrt{n}}\int_0^C h(w)dw\right)\right) \\ \to\ & 0, \end{aligned}$$

as $n \to \infty$, because $a^n \Rightarrow e$ as $n \to \infty$ from (141) and

$$\sqrt{n}\left(1 - \exp\left(\frac{-1}{\sqrt{n}}\int_0^C h(w)dw\right)\right) \to \int_0^C h(w)dw < \infty,$$

as $n \to \infty$. We conclude from (227), (229), (230), and (231) that

$$n^{-1/2} \sum_{i=A^n \circ a^n(\cdot)}^{A^n(\cdot)} \mathbf{1}\left\{V^n\left(t_i^{n,-}\right) \geq a_i^n\right\} \Rightarrow 0,$$

as $n \to \infty$, also using (141) and the fact that $\overline{A}^n \to e$ as $n \to \infty$, almost surely, uniformly on compact sets. □

Proof of Lemma 3.5.4: For each $x \in \Re$, let P_x be a probability measure such that the Brownian motion W has initial position $W(0) = x$. We first argue that for $0 \leq x \leq C$, $P_x\left(T_0^C < \infty\right) = 1$. Define

$$\tilde{T}_0^C \equiv \inf\{t \geq 0 : \phi_C(W)(t) = 0\},$$

and observe that

$$P_x\left(\tilde{T}_0^C < \infty\right) = 1 \tag{232}$$

by Problem 7 in Chapter 5 in Harrison [17]. From (109), for any $x \in D([0,\infty), \Re)$,

$$x(t) = \mathcal{M}_C^h(x)(t) + \int_0^t \left(\int_0^{\phi_C\left(\mathcal{M}^h(x)\right)(s)} h(u)du\right) ds$$

is written as the sum of $\mathcal{M}_C^h(x)$ and a non-decreasing function. Then, Theorem 1.6 in [29] establishes

$$\phi_C\left(\mathcal{M}_C^h(x)\right) \leq \phi_C(x)$$

for any $x \in D\left([0,\infty), \Re\right)$. We conclude $T_0^C \leq \tilde{T}_0^C$ on every sample path, which from (232) implies $P_x\left(T_0^C < \infty\right) = 1$.

To see $P_x\left(T_0 < \infty\right) = 1$ for all $x \geq 0$, first observe that for

$$A \equiv \frac{\sigma^2}{2}\frac{d^2}{dx^2} + \left(\theta - H(x)\right)\frac{d}{dx},$$

the function

$$u_n^\epsilon(x) = 1 - \frac{\int_\epsilon^x \exp\left(\frac{2}{\sigma^2}\int_0^y \left(H(z) - \theta\right) dz\right) dy}{\int_\epsilon^n \exp\left(\frac{2}{\sigma^2}\int_0^y \left(H(z) - \theta\right) dz\right) dy} \tag{233}$$

solves the ordinary differential equation

$$\left(Au_n^\epsilon\right)(x) = 0,\ \epsilon \leq x \leq n, \tag{234}$$

with boundary conditions

$$u_n^\epsilon(\epsilon) = 1,\ u_n^\epsilon(n) = 0. \tag{235}$$

Next consider the diffusion

$$Z(t) = -\int_0^t H\left(Z(s)\right) ds + W(t),$$

and observe that if

$$\begin{aligned} T_\epsilon &\equiv \inf\left\{t \geq 0 : \phi^h(W)(t) \leq \epsilon\right\} \\ T_n &\equiv \inf\left\{t \geq 0 : \phi^h(W)(t) \geq n\right\} \\ T_\epsilon^z &\equiv \inf\left\{t \geq 0 : Z(t) \leq \epsilon\right\} \\ T_n^z &\equiv \inf\left\{t \geq 0 : Z(t) \geq n\right\}, \end{aligned}$$

then for $\epsilon < x < n$,

$$P\left(T_\epsilon < T_n | W(0) = x\right) = P_x\left(T_\epsilon^z < T_n^z | W(0) = x\right), \tag{236}$$

because for $0 \leq t < T_\epsilon^z \wedge T_n^z$, $Z(t) \in (\epsilon, n)$, and so $\phi^h(W)(t) = Z(t)$. Let $\tilde{u}_n^\epsilon$ be a bounded, twice continuously differentiable function such that

$$\tilde{u}_n^\epsilon(x) = u_n^\epsilon(x) \text{ for all } \epsilon \leq x \leq n,$$

and having bounded first derivative. Ito's formula establishes

$$d\tilde{u}_n^\epsilon\left(Z(t)\right) dt = \left(A\tilde{u}_n^\epsilon\right)\left(Z(t)\right) dt + \tilde{u}_n^{\epsilon'}\left(Z(t)\right) \sigma dW(t). \tag{237}$$

Our assumption that $\tilde{u}_n^{\epsilon'}$ is bounded is a sufficient condition for the stochastic integral in (237) to be a martingale. Applying the optional stopping theorem to the bounded stopping time $t \wedge T_\epsilon^z \wedge T_n^z$ and using the fact that $\left(A\tilde{u}_n^\epsilon\right)\left(Z(t)\right) = 0$ for $0 \leq t < T_\epsilon^z \wedge T_n^z$ from (234), we find

$$\tilde{u}_n^\epsilon(x) = E\left[\tilde{u}_n^\epsilon\left(Z\left(t \wedge T_\epsilon^z \wedge T_n^z\right)\right) | W(0) = x\right]. \tag{238}$$

The exit time of a one-dimensional diffusion from a compact subinterval of $\Re$ is finite with probability 1. (See Section 5 in [23].) Therefore, taking the limit as $t \to \infty$ in (238) and using the bounded convergence theorem shows

$$\tilde{u}_n^\epsilon(x) = E_x\left[\tilde{u}_n^\epsilon\left(Z\left(T_\epsilon^z \wedge T_n^z\right)\right)\right].$$

The boundary conditions in (235) and the equality in (236) then imply

$$\tilde{u}_n^\epsilon(x) = P_x\left(T_\epsilon^z < T_n^z\right) = P_x\left(T_\epsilon < T_n\right). \tag{239}$$

Since the diffusion $\phi^h(W)$ has continuous paths almost surely,

$$\lim_{n\to\infty} \lim_{\epsilon \downarrow 0} P_x\left(T_\epsilon < T_n\right) = P_x\left(T_0 < \infty\right). \tag{240}$$

Furthermore, for any $x \geq 0$, from the expression for u_n^ϵ in (233),

$$\lim_{n\to\infty} \lim_{\epsilon \downarrow 0} \tilde{u}_n^\epsilon(x) = 1, \tag{241}$$

because

$$\int_0^n \exp\left(\frac{2}{\sigma^2}\int_0^y \left(H(z)-\theta\right) dz\right) dy \to \infty,$$

as $n \to \infty$, under the assumption that $\lim_{z\to\infty} H(z) > \theta$. We conclude from (239), (240), and (241) that

$$1 = P_x\left(T_0 < \infty\right).$$

□

REFERENCES

[1] U. B. of Labor Statistics, Table b-1: Employees on nonfarm payrolls by major industry, 1950 to date. As reported on www.bls.gov.

[2] Ata, B., Harrison, J., and Shepp, L., "Drift rate control of a Brownian processing system," *Annals of Applied Probability*, vol. 15, no. 2, pp. 1145–1160, 2005.

[3] Baccelli, F., Boyer, P., and Hebuterne, G., "Single-server queues with impatient customers," *Advances in Applied Probability*, vol. 16, pp. 887–905, 1984.

[4] Billingsley, P., *Convergence of Probability Measures*. New York: John Wiley & Sons, Inc., 1999. Second Edition.

[5] Billingsley, P., *Convergence of Probability Measures*. New York: John Wiley and Sons, 1999.

[6] Borovkov, A., "On limit laws for service processes in multi-channel systems," *Siberian Journal of Mathematics*, vol. 8, pp. 983–1004, 1967.

[7] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L., "Statistical analysis of a telephone call center: A queueing-science perspective," *Journal of the American Statistical Association*, vol. 100, pp. 36–50, 2005.

[8] Chen, H. and Yao, D. D., *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. New York: Springer-Verlag, 2001.

[9] Dai, J. G. and Dai, W., "A heavy traffic limit theorem for a class of open queueing networks with finite buffers," *Queueing Systems*, vol. 32, pp. 5–40, 1999.

[10] Doytchinov, B., Lehoczky, J., and Shreve, S., "Real-time queues in heavy traffic with earliest-deadline-first queue discipline," *Annals of Applied Probability*, vol. 11, no. 2, pp. 332–378, 2001.

[11] Echeverria, P., "A criterion for invariant measures of Markov processes," *Z. Wahrsch. Verw. Gebiete*, vol. 61, pp. 1–16, 1982.

[12] Evans, L. C. and Gariepy, R. F., *Measure Theory and Fine Properties of Functions*. Boca Raton: John Wiley & Sons, 1992.

[13] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers: Tutorial, review and research prospects," *Manufacturing and Service Operations Management*, vol. 5, pp. 79–141, 2003.

[14] Garnett, O., Mandelbaum, A., and Reiman, M. I., "Designing a call center with impatient customers," *Manufacturing and Service Operations Management*, vol. 4, pp. 208–227, 2002.

[15] Halfin, S. and Whitt, W., "Heavy taffic limits for queues with many exponential servers," *Operations Research*, vol. 29, pp. 567–588, 1981.

[16] Hall, P. and Heyde, C. C., *Martingale Limit Theory and Its Applications*. New York: Academic Press, 1980.

[17] Harrison, J. M., *Brownian Motion and Stochastic Flow Systems*. Malabar: Krieger Publishing Company, 1985.

[18] Harrison, J. M. and Reiman, M. I., "Reflected Brownian motion on an orthant," *Annals of Probability*, vol. 9, no. 2, pp. 302–308, 1981.

[19] Harrison, J. M. and Williams, R. J., "Multidimensional reflected Brownian motions having exponential stationary distributions," *Annals of Probability*, vol. 15, pp. 115–137, 1987.

[20] Iglehart, D. L. and Whitt, W., "Multiple channel queues in heavy traffic, I and II," *Advances in Applied Probability*, vol. 2, pp. 150–177 and 355–364, 1970.

[21] Jacod, J. and Shiryaev, A., *Limit Theorems for Stochastic Processes*. New York: Springer, 2003.

[22] Jelenkovic, P., Mandelbaum, A., and Momcilovic, P., "Heavy traffic limits for queues with many determinsitic servers," *Queueing Systems*, vol. 47, pp. 53–69, 2005.

[23] Karatzas, I. and Shreve, S. E., *Brownian Motion and Stochastic Calculus*. New York: Springer, 1991. Second Edition.

[24] Karlin, S. and Taylor, H. M., *A First Course in Stochastic Processes*. New York: Academic Press, 1975.

[25] Kingman, J. F. C., "The single server queue in heavy traffic," *Proc. Cambridge Philos. Soc.*, vol. 57, pp. 902–904, 1961.

[26] Kingman, J. F. C., "Two similar queues in parallel," *Ann. Math. Statist.*, vol. 32, pp. 1314–1323, 1961.

[27] Kingman, J. F. C., "On queues in heavy traffic," *J. Roy. Statist. Soc. Ser. B*, vol. 24, pp. 383–392, 1962.

[28] Krichagina, E. and Puhalskii, A., "A heavy traffic analysis of a closed queueing system with a GI/∞ service center," *Queueing Systems*, vol. 25, pp. 235–280, 1997.

[29] Kruk, L., Lehoczky, J., Ramanan, K., and Shreve, S., "An explicit formula for double reflected processes in [0,a]," 2005. Submitted.

[30] Kruk, L., Lehoczky, J., and Shreve, S., "Accuracy of state space collapse for earliest-deadline-first queues," *Annals of Applied Probability*, vol. 16, no. 2, pp. 516–581, 2006.

[31] Kruk, L., Lehoczky, J., Shreve, S., and Yeung, S., "Multiple-input heavy-traffic real-time queues," *Annals of Applied Probability*, vol. 13, no. 1, pp. 54–99, 2003.

[32] Kruk, L., Lehoczky, J., Shreve, S., and Yeung, S., "Earliest-deadline-first service in heavy traffic acyclic networks," *Annals of Applied Probability*, vol. 14, no. 3, pp. 1306–1352, 2004.

[33] Kulkarni, V. G., M*odeling,* A*nalysis,* D*esign, and* C*ontrol of* S*tochastic* S*ystems*. Springer, 1989.

[34] Lehoczky, J., "Real-time queueing theory," in *Proceedings of the IEEE Real-Time Systems Symposium*, pp. 186–195, IEEE, 1997.

[35] Lillo, R. and Martin, M., "Stability in queues with impatient customers," *Stochastic Models*, vol. 17, 2001.

[36] Liptser, R. S. and Shiryaev, A., *Theory of* M*artingales*. Kulwer, 1989.

[37] Mandelbaum, A. and Momcilovic, P., "Queues with many servers: The virtual waiting-time process in the QED regime," *Submitted to Mathematics of Operations Research*, 2005.

[38] Mandelbaum, A. and Shimkin, N., "A model for rational abandonment from invisible queues," *Queueing Systems*, vol. 36, pp. 141–173, 2000.

[39] Palm, C., "Etude des delais d'attente," *Ericsson Technics*, vol. 5, pp. 37–56, 1937.

[40] Puhalskii, A. and Reiman, M. I., "The multiclass GI/PH/N queue in the haflin-whitt regime," *Advances in Applied Probability*, vol. 32, pp. 564–595, 2000.

[41] Reed, J. E., "The G/GI/N queue in the Halfin-Whitt regime: Idle time equations," *In Preparation*.

[42] Reed, J. E. and Ward, A. R., "A diffusion approximation for a generalized jackson network with reneging," in *Proceedings of the 42nd Allerton Conference on Communication, Control, and Computing.*, 2004.

[43] Reiman, M. I., "The heavy traffic diffusion approximation for sojourn times in Jackson networks," in *Applied Probability-Computer Science, The Interface* (Disney, R. L. and Ott, T. J., eds.), Boston: Birkhauser, 1982.

[44] Reiman, M. I., "Some diffusion approximations with state space collapse," in *Modelling and Performance Evaluation Methodology* (Baccelli, F. and Fayolle, G., eds.), pp. 209–240, Springer-Verlag, 1984.

[45] Resnick, S. I., *A Probability Path*. Boston: Birkhauser, 1999.

[46] Skorokhod, A. V., "Stochastic equations for diffusions in a bounded region 1,2," *Theor. of Prob. and Its Appl.*, vol. 6, pp. 264–274, 1961.

[47] Stanford, R. E., "Reneging phenomena in single channel queues," *Mathematics of Operations Research*, vol. 4, pp. 162–178, 1979.

[48] Utchitelle, L., "Answering '800' calls, extra income buy no security," *The New York Times*. March 27, Section A, p.1. Col 5, 2002.

[49] Ward, A. R. and Bambos, N., "On stability of queueing networks with job deadlines," *Journal of Applied Probability*, vol. 40, pp. 293–304, 2003.

[50] Ward, A. R. and Glynn, P., "A diffusion approximation for a markovian queue with reneging," *Queueing Systems*, vol. 44, pp. 109–123, 2003.

[51] Ward, A. R. and Glynn, P., "Properties of the reflected ornstein-uhlenbeck process," *Queueing Systems*, vol. 44, pp. 109–123, 2003.

[52] Ward, A. R. and Glynn, P., "A diffusion approximation for a GI/GI/1 queue with balking or reneging," *Queueing Systems*, vol. 50, pp. 371–400, 2005.

[53] Ward, A. R. and Kumar, S., "Asymptotically optimal admission control of a queue with impatient customers," *Revised for Mathematics of Operations Research*, 2006.

[54] Whitt, W., *Stochastic-Process Limits*. New York: Springer, 2002.

[55] Whitt, W., "Heavy-traffic limits for loss proportions in single-server queues," *Queueing Systems*, vol. 46, pp. 507–536, 2004.

[56] Whitt, W., "Engineering solution of a basic call center model," *Management Science*, vol. 51, no. 2, pp. 221–235, 2005.

[57] Whitt, W., "Heavy-traffic limits for the G/H2/n/m queue," *Mathematics of Opertions Research*, vol. 30, pp. 1–27, 2005.

[58] Whitt, W., "Fluid models for multi-server queues with abandonments," *Operations Research*, vol. 54, pp. 37–54, 2006.

[59] Zeltyn, S. and Mandelbaum, A., "Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue," *Queueing Systems*, vol. 51, pp. 361–402, 2005.

[60] Zohar, E., Mandelbaum, A., and Shimkin, N., "Adaptive behavior of impatient customers in tele-queues: Theory and empirical support," *Management Science*, vol. 48, pp. 566–583, 2002.